

# Mutual Information-Based Visual Servoing

Amaury Dame and Eric Marchand

**Abstract**—In this paper, we propose a new information theoretic approach to achieve visual servoing directly utilizing the information (as defined by Shannon) contained in the images. A metric derived from information theory, i.e., mutual information, is considered. Mutual information is widely used in multimodal image registration since it is insensitive to changes in the lighting condition and to a wide class of nonlinear image transformations. In this paper, mutual information is used as a new visual feature for visual servoing, which allows us to build a new control law that can control the six degrees of freedom (DOF) of a robot. Among various advantages, this approach requires no matching or tracking step, is robust to large illumination variations, and allows the consideration of different image modalities within the same task. Experiments on a real robot demonstrate the efficiency of the proposed visual-servoing approach.

**Index Terms**—Entropy, mutual information (MI), visual servoing.

## I. INTRODUCTION

VISUAL servoing uses the information provided by a vision sensor to control the movements of a dynamic system [3], [4]. This approach requires the extraction of visual information (usually geometric features) from the image in order to design the control law. Robust extraction and real-time spatiotemporal tracking of these visual cues [18] is a nontrivial task and is one of the bottlenecks of the expansion of visual servoing.

Recently, it has been shown that no information other than the image intensities can be considered to control the robot motion and that the classical tracking and matching processes can be avoided. The approaches proposed by Collewet and Marchand [5], Deguchi [8], and Kallem *et al.* [12] no longer require any matching or tracking processes. Assuming that the pixel intensities at the desired pose are known and considering the whole set of image intensities as a feature avoid the tracking and matching processes. Following this, various approaches

have been presented. Deguchi [8] and Nayar *et al.* [19] consider the full image, but in order to reduce the dimensionality of image data, they perform an eigenspace decomposition of the image. The control is then performed directly in the eigenspace requiring both an off-line computation of this eigenspace (using a principal component analysis) and the projection of each acquired image on this subspace. To create a control law closer to the image, Collewet and Marchand [5] propose to regulate directly the sum of squared differences (SSD) between the current and reference images. Such an approach is nevertheless quite sensitive to illumination variations (although using a more complex illumination model in some particular cases is possible). Kallem *et al.* [12] also consider the pixel intensities with a kernel-based method that leads to a highly decoupled control law. However, this approach cannot control the six degrees of freedom (DOF) of the robot, and it is very limited in the case of appearance variations. Another approach that does not require tracking or matching has been proposed in [1]. It models collectively feature points extracted from the image as a mixture of Gaussian and attempts to minimize the distance function between the Gaussian mixture at current and desired poses. Simulation results show that this approach is able to control the three DOF of the robot. However, an image processing step is still required to extract the current feature points.

Although these methods are very different from the classical geometric approaches [3], the goal remains the same: From its current pose  $\mathbf{r}$ , the robot has to reach the desired pose  $\mathbf{r}^*$ . In terms of optimization [17], it means that during the whole visual-servoing task, the pose of the robot has to evolve in the direction of a given alignment function extremum. The camera velocity is then computed using the derivatives of the cost function with respect to the pose  $\mathbf{r}$ .

As previously stated, image intensities are quite sensitive to modifications of the environment [5]. To solve this problem, our new approach does not consider directly the luminance of the pixels but the information contained in the images. The visual feature is the mutual information (MI) defined by Shannon in [24]. The MI (built from the image entropy) of two random variables (images) measures their mutual dependence. This function does not directly compare the intensities of the two images but the distribution of the information in the images. Given the two images, the higher the MI, the better the alignment between the two images. To consider the information contained in the image and not the image itself offers a measure robust to perturbations or the image modalities (as soon as enough information is shared between the modalities). This yields very interesting properties for visual servoing: As for [5], this approach does not require any tracking or matching step, it is robust to large illumination variations and to partial occlusions and is able to consider different image modalities in the acquisition process. Although MI has been widely used for

Manuscript received July 15, 2010; revised January 21, 2011; accepted April 19, 2011. Date of publication May 23, 2011; date of current version October 6, 2011. This paper was recommended for publication by Associate Editor K. Kyriakopoulos and Editor G. Oriolo upon evaluation of the reviewers' comments. This work was supported by La Délégation Générale pour l'Armement under contribution to student grant. Part of this paper was published in the Proceedings of the International Conference on Robotics and Automation (ICRA) 2009 [6] and ICRA 2010 [7].

A. Dame is with the Lagadic Team, CNRS, INRIA Rennes—Bretagne Atlantique, IRISA, Rennes 35000, France (e-mail: amaury.dame@irisa.fr).

E. Marchand is with the Lagadic Team, Université de Rennes 1, IRISA, INRIA Rennes—Bretagne Atlantique, Rennes 35042, France (e-mail: marchand@irisa.fr).

This paper has supplementary downloadable material available at <http://ieeexplore.ieee.org>.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TRO.2011.2147090

multimodal medical image registration [27] and more recently in tracking [10], to the best of our knowledge, this is the first time that it has been considered to build a vision-based control law.

The remainder of this paper is organized as follows. In Section II, we recall that visual servoing can be easily formulated as an optimization problem and recall the main differences between feature-based visual-servoing and direct visual-servoing approaches. Section III gives a background on information theory and describes the new metric based on MI related to images. The variation of the MI with respect to the displacement of the camera is defined in Section IV, and the resulting control law, which is based on the optimization of MI, is presented in Section V. Finally, some positioning and navigation tasks of a 6-DOF robot are presented in Section V, and a more general validation of the proposed approach is given by an empirical convergence analysis in Section VI.

## II. FROM FEATURE-BASED VISUAL SERVOING TO DIRECT APPROACHES

For years, image-based visual servoing (IBVS) has been mostly known through feature-based approaches [3]. More recently, keeping the formulation of the positioning task as an optimization problem, direct visual-servoing approaches [5], [8], [12] have been proposed. These approaches have the advantage that they do not require any feature extraction, matching, and tracking steps; therefore, they are very accurate. Within this class of methods, we propose in this paper a new information theoretic approach that redefines the camera alignment process using the Shannon MI.

### A. Visual Servoing as an Optimization Approach

A visual-servoing problem can always be written as an optimization problem [17]. The goal of visual servoing is that, from an initial arbitrary pose, the camera pose  $\mathbf{r}$  reaches the desired pose  $\mathbf{r}^*$  that best satisfies some properties measured in or from the images. If we note  $f$ , i.e., the function that measures the positioning error, then the visual-servoing task can be written as

$$\hat{\mathbf{r}} = \arg \min_{\mathbf{r}} f(\mathbf{r}, \mathbf{r}^*) \quad (1)$$

where  $\hat{\mathbf{r}}$  is the camera pose obtained after the visual-servoing task. The visual-servoing problem can, therefore, be considered as an optimization of the function  $f$  where  $\mathbf{r}$  is incrementally updated to reach an optimum of  $f$  at  $\hat{\mathbf{r}}$ . If  $f$  is correctly chosen at the end of the minimization, the final camera pose  $\hat{\mathbf{r}}$  should be equal to the desired one  $\mathbf{r}^*$ . For an eye-in-hand configuration, the pose update is performed by applying a velocity  $\mathbf{v}$ , which corresponds to the direction of the alignment function descent, to the camera that is mounted on a robot end-effector:

$$\mathbf{r}_{k+1} = \mathbf{r}_k \oplus \mathbf{v} \quad (2)$$

where “ $\oplus$ ” is the operator that updates the pose and which is “implemented” through the robot controller.

### B. Feature-Based Approaches

Classical visual-servoing approaches consider a function  $f$  based on the distance between geometrical features extracted from the image. The visual features  $\mathbf{s}$  can be 2-D features that lead to an IBVS approach or 3-D features (such as the camera pose) that lead to a position-based visual-servoing approach. These visual features (points, lines, moments, contours, pose, etc.) have thus to be selected and extracted from the images to control the desired DOF of the robot. The control law is then designed so that these visual features  $\mathbf{s}(\mathbf{r})$  reach a desired value  $\mathbf{s}^*$ , which leads to a correct realization of the task. The optimization problem can thus be written as

$$\hat{\mathbf{r}} = \arg \min_{\mathbf{r}} \|\mathbf{s}(\mathbf{r}) - \mathbf{s}^*\|. \quad (3)$$

Although very efficient, these approaches have some drawbacks. First, some features have to be chosen depending on the scene characteristics. Second, the current features  $\mathbf{s}(\mathbf{r})$  have to be tracked in real time and matched with the desired ones  $\mathbf{s}^*$ . Despite the recent advances in computer vision, the tracking issue is far from being solved. Finally, these tracking and matching tasks are prone to some measurement errors that cause the visual-servoing task to be less accurate than it could be.

### C. Direct Approaches

To avoid these issues inherent to the use of geometrical features, other formulations that use the images as a whole have been proposed. We refer to this class of methods as direct visual-servoing approaches. In this context, the visual-servoing task is defined as an alignment between the current image  $\mathbf{I}(\mathbf{r})$  and the image acquired at the desired camera pose  $\mathbf{I}^*$ . The camera is controlled in order to minimize an error measured between the current and desired images.

1) *Kernel and Photometric Visual Servoing*: One solution has been to consider a kernel-based approach [12]. This approach shows a large convergence domain; nevertheless, it gives no precise alignment information and the visual-servoing task is limited to 4 DOF. Furthermore, it is very sensitive to illumination variations.

Another solution, i.e., the photometric visual-servoing approach, considers  $f$  as the SSD of the image intensities [5]. In this case, the optimization can simply be written as

$$\begin{aligned} \hat{\mathbf{r}} &= \arg \min_{\mathbf{r}} \|\mathbf{I}(\mathbf{r}) - \mathbf{I}^*\| \\ &= \arg \min_{\mathbf{r}} \sum_{\mathbf{x}} (\mathbf{I}(\mathbf{r}, \mathbf{x}) - \mathbf{I}^*(\mathbf{x}))^2 \end{aligned} \quad (4)$$

where  $\mathbf{I}(\mathbf{r}, \mathbf{x})$  is the intensity of the pixel  $\mathbf{x}$  in the image  $\mathbf{I}$  acquired at the current pose  $\mathbf{r}$ . That equation is, in fact, a reformulation of (3), where the feature vector  $\mathbf{s}$  is defined by the image intensities. As already stated, the main advantage of these direct visual-servoing approaches is that they do not rely on any tracking or matching process. Furthermore, since the feature vector contains all the image information and since no intermediate visual features are used, the resulting visual-servoing process does not suffer from measurement errors and performs



Fig. 1. From feature-based tracking to direct approaches. A visual-servoing task was usually based on measures extracted from the image such as points and contours. It is now possible to run a visual-servoing task with no feature extraction or image processing through direct approaches. Our new approach is now capable of servoing the robot using images acquired using different modalities.

a very accurate positioning task under constant illumination conditions.

One problem remains in these solutions: If the appearance of the object has changed from  $\mathbf{I}^*$  to the current acquired image  $\mathbf{I}(\mathbf{r})$  due to some illumination variations or some occlusions, then the cost function is highly affected, which causes the visual-servoing task to diverge.

2) *Proposed Approach*: The solution, which is proposed in this paper, is to define the alignment function  $f$  as the MI between the two images. The MI can be defined as the quantity of information shared by two signals (or images in our case). This metric is very robust to the appearance variations. As in the other direct approaches, we still use the entire information provided by the images  $\mathbf{I}(\mathbf{r})$  and  $\mathbf{I}^*$  by achieving the following optimization:

$$\hat{\mathbf{r}} = \arg \max_{\mathbf{r}} \text{MI}(\mathbf{I}(\mathbf{r}), \mathbf{I}^*). \quad (5)$$

In the following section, we will see that the optimization of MI is well adapted for the visual-servoing problem: It performs a very accurate positioning task, has a large convergence area, and is robust to both occlusions and illumination variations (see Fig. 1).

Finally, it opens new possible visual-servoing applications. Indeed, it is also robust to the alignment between images acquired using different sensors (modalities). For example, Fig. 1 shows a map and a satellite image of the same area that are used in the same visual-servoing task.

### III. INFORMATION THEORY AND MUTUAL INFORMATION

In this section, a brief definition of MI is given as it was originally defined in the theory of communication in [24]. A definition adapted to the optimization of the visual-servoing problem is then derived from the original definition to best fit the optimization problem.

#### A. Shannon Mutual Information

MI is an alignment function that was first introduced in information theory. Some essential notions such as entropy and joint entropy underpin the application of this alignment measure. To address these definitions, let us omit the pose  $\mathbf{r}$  for the purpose of clarity and 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  $\mathbf{x}$ ).

1) *Entropy*: The entropy  $H(\mathbf{I})$  is a measure of variability of a random variable  $\mathbf{I}$ . If  $i$  is a possible value of  $\mathbf{I}(\mathbf{x})$  ( $i \in [0, N_{c_I}]$  with  $N_{c_I} = 255$ ) 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:

$$H(\mathbf{I}) = - \sum_{i=0}^{N_{c_I}} p_{\mathbf{I}}(i) \log(p_{\mathbf{I}}(i)). \quad (6)$$

The log basis only changes the unit of the entropy; therefore, it makes no difference in our optimization problem. The formulation can be seen as follows: Since  $-\log(p_{\mathbf{I}}(i))$  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.

2) *Joint Entropy*: Following the same principle, the 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

$$H(\mathbf{I}, \mathbf{I}^*) = - \sum_{i=0}^{N_{c_I}} \sum_{j=0}^{N_{c_{I^*}}} p_{\mathbf{II}^*}(i, j) \log(p_{\mathbf{II}^*}(i, j)) \quad (7)$$

where  $i$  and  $j$  are, respectively, the possible values of the variables  $\mathbf{I}$  and  $\mathbf{I}^*$ , and  $p_{\mathbf{II}^*}(i, j) = \Pr(\mathbf{I}(\mathbf{x}) = i \cap \mathbf{I}^*(\mathbf{x}) = j)$  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  $\mathbf{I}^*$ .

At first sight, the joint entropy could be considered as a good alignment measure: If the dispersion of the joint histogram is small, then the correlation between the two images is strong and we can suppose that the two images are aligned. Nevertheless, the dependences on the entropies of  $\mathbf{I}$  and  $\mathbf{I}^*$  make it unsuitable. Indeed, if one of the images has a constant gray-level value, then the joint histogram would be very focused and the entropy value would be very small, despite the fact that the two images are not aligned.

3) *Original Mutual Information*: The definition of MI solves the aforementioned problem [24], [27]. To subtract the random

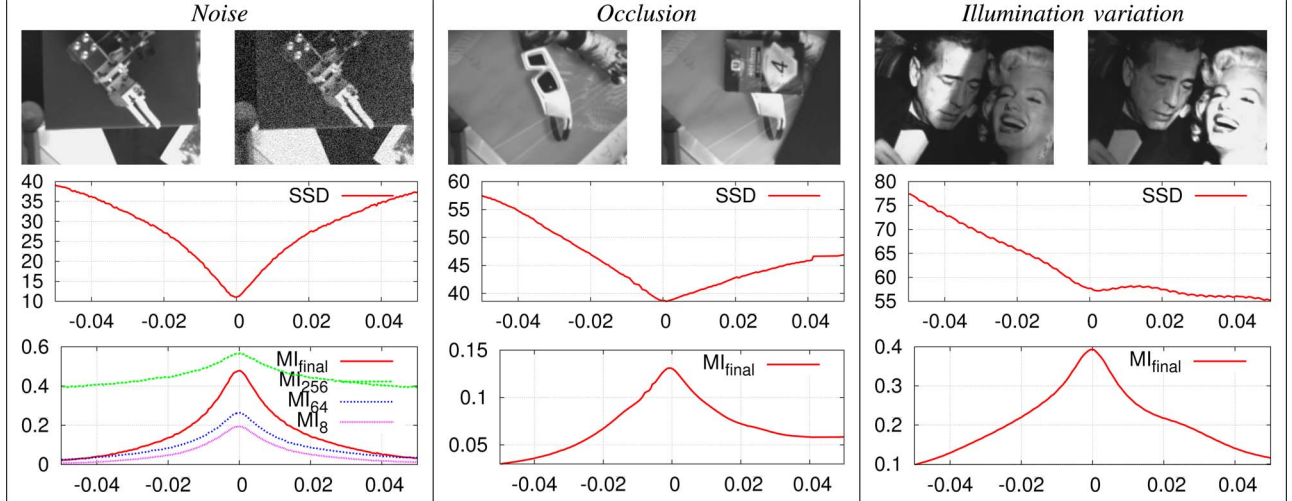


Fig. 2. MI robustness with respect to noise, occlusions, and illumination variations. The second and third lines show, respectively, the values of the SSD and MI alignment functions between the desired and current images with respect to the camera translational error (m). MI is robust in every case, whereas the SSD is not in the illumination case. The noisy case depicts the improvement made on the MI computation from its original version  $MI_{256}$  to the one used in the visual-servoing approach  $MI_{\text{final}}$ .

variable's entropies from their joint entropy yields an alignment measure that does not depend on the variable marginal entropies. The MI of two random variables  $\mathbf{I}$  and  $\mathbf{I}^*$  is then given by

$$MI(\mathbf{I}, \mathbf{I}^*) = H(\mathbf{I}) + H(\mathbf{I}^*) - H(\mathbf{I}, \mathbf{I}^*) \quad (8)$$

where MI measures the quantity of information shared by two random variables.

4) *Link With Visual Servoing*: If this expression is combined with the previously defined visual-servoing problem, we can consider that the image or random variable  $\mathbf{I}$  depends on the pose of the camera  $\mathbf{r}$ . Using the same notations as in Section II, the MI can thus be written with respect to  $\mathbf{r}$  as follows:

$$MI(\mathbf{r}) = MI(\mathbf{I}(\mathbf{r}), \mathbf{I}^*) = H(\mathbf{I}(\mathbf{r})) + H(\mathbf{I}^*) - H(\mathbf{I}(\mathbf{r}), \mathbf{I}^*). \quad (9)$$

To develop the MI expression, we assume that the histogram of the current image and the joint histogram of the two images also depend on the camera pose. The probabilities  $p_{\mathbf{I}}$  and  $p_{\mathbf{II}^*}$  are thus noted with respect to the current pose. The MI between the current and desired images can, therefore, be rewritten as

$$MI(\mathbf{r}) = \sum_{i,j} p_{\mathbf{II}^*}(i, j, \mathbf{r}) \log \left( \frac{p_{\mathbf{II}^*}(i, j, \mathbf{r})}{p_{\mathbf{I}}(i, \mathbf{r})p_{\mathbf{I}^*}(j)} \right). \quad (10)$$

To illustrate the original MI function in the positioning problem, Fig. 2 (green curve  $MI_{256}$ ) shows the results obtained in a simple example where the displacement of the camera is limited to one translation along the  $x$ -axis of the camera frame (that is,  $\mathbf{r} = t_x$ ). The desired image  $\mathbf{I}^*$  is acquired and then the camera is moved around the desired pose with respect to the translation  $t_x$  where the current images  $\mathbf{I}(\mathbf{r})$  are acquired. To check the robustness of MI with respect to noise, a Gaussian white noise is added to each pixel intensity and  $MI(\mathbf{r})$  is computed for each pose. Fig. 2 represents the corresponding values of the MI and the SSD with respect to the positioning error  $\Delta \mathbf{r}$ .

The computation of MI using the original approach yields an accurate cost function, but it has two problems: First, the computation of a classical histogram is not differentiable, and it is

also sensitive to small local maxima. Indeed the pixel intensities of the numerical images are encoded on 256 gray-level values. In that case, the histograms and joint histograms have, respectively, 256 and  $256 \times 256$  bins. To consider such a number of bins implies that several histogram bins are empty. Perturbations on these kinds of histograms have then a strong impact on the entropy measures.

### B. Adapting the Mutual Information Formulation

The original definition of MI requires the computation of large histograms that are highly time consuming, not differentiable, and yields local maxima. Therefore, the original formulation is not adapted for our gradient-based optimization problem and requires modifications.

1) *Histograms Binning*: Starting from the previous observation, one obvious solution is to decrease the number of histogram bins [21]. The analytical formulation of the normalized histogram of an image  $\mathbf{I}$  is generally written as

$$p_{\mathbf{I}}(i, \mathbf{r}) = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \phi(i - \mathbf{I}(\mathbf{r}, \mathbf{x})) \quad (11)$$

where  $\mathbf{x}$  are the pixels of the image, and  $N_{\mathbf{x}}$  is the number of pixels. Each time  $\mathbf{I}(\mathbf{r}, \mathbf{x}) = i$ , the  $i$ th histogram bin entry is typically incremented by 1.  $\phi$  is then a Kronecker's delta function defined by  $\phi(i - i') = \delta_{ii'} = 1$  for  $i = i'$  and  $\phi(i - i') = 0$  otherwise.

As can be seen, the number of bins corresponds to the maximum gray-level intensity of the image  $N_{c1} = 255$ . To reduce it, the image intensities are simply scaled as follows:

$$\bar{\mathbf{I}}(\mathbf{r}, \mathbf{x}) = \mathbf{I}(\mathbf{r}, \mathbf{x}) \frac{N_c}{N_{c1}} \quad (12)$$

where  $N_c$  is the new number of histogram bins. The obtained intensities are no longer integer values. A classical method would then be to simply use the integer part of  $\bar{\mathbf{I}}$  to compute the new

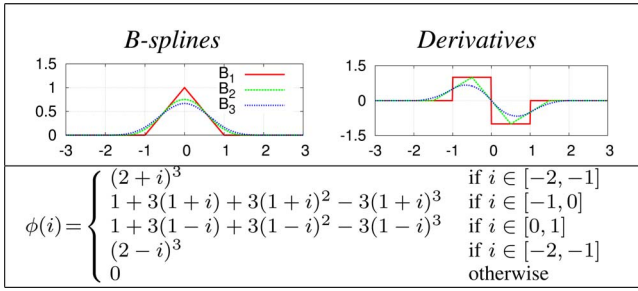


Fig. 3. B-splines functions used for an efficient and differentiable computation of the histograms and their derivatives.

histogram. However, this solution is not suitable for two reasons: The computation is still not differentiable, and the loss of the decimal part involves a large loss of the information provided by the intensity.

An adequate solution is to keep the real value  $\bar{I}$ . Instead of incrementing one entry of the histogram for each pixel, several entries are incremented, depending on their distance with the input intensity  $\bar{I}$ . To do so, Viola and Wells [27] introduced the use of Gaussian density functions for  $\phi$ , while Maes *et al.* [15] used partial volume interpolation by choosing  $\phi$  as B-spline functions that are typically an approximation of Gaussian functions. In this paper, we focus on the use of B-spline functions (see Fig. 3) for their advantage concerning the computation time. Moreover, their properties are well adapted to histogram computation and optimization problems: Their use in the histogram computation requires no renormalization, and their derivatives are easily and inexpensively computed.

Using both the scaled images and the B-spline function  $\phi$ , the computation of the probabilities and joint probability used in (6) and (7) are written as

$$\begin{aligned} p_{\mathbf{I}}(i, \mathbf{r}) &= \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \phi(i - \bar{\mathbf{I}}(\mathbf{r}, \mathbf{x})) \\ p_{\mathbf{I}^*}(j) &= \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \phi(j - \bar{\mathbf{I}}^*(\mathbf{x})) \\ p_{\mathbf{II}^*}(i, j, \mathbf{r}) &= \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \phi(i - \bar{\mathbf{I}}(\mathbf{r}, \mathbf{x})) \phi(j - \bar{\mathbf{I}}^*(\mathbf{x})). \end{aligned} \quad (13)$$

Several solutions have been proposed to estimate an optimal number of histogram bins [23], [25]. Nevertheless, a constant number of bins set to  $N_c = 8$ , which keeps a small value and avoids losing the information, has always given satisfactory results in our experiments. Thus, it will be the solution adopted in the remainder of this paper.

If we compare in Fig. 2 the MI values between the original formulation (see the  $\text{MI}_{256}$  curve) and the one with 8 ( $\text{MI}_8$ ) and 64 bins ( $\text{MI}_{64}$ ) in the histograms, the benefits of the histogram binning operation are obvious. MI is no longer subject to local maxima, and as the number of bins decreases, the function is greatly convexified. In terms of optimization, the convergence domain is then greatly widened.

2) *Image Filtering*: The histogram binning operation gives a very satisfying MI function. Nevertheless, some works on reg-

istration by MI maximization have shown that the convergence domain can be increased using some particular image interpolation [21], that is, the process necessary to pick up an intensity at a noninteger position in the image.

In this approach, image intensities are always picked at integer positions; thus, no image interpolation is originally required. To choose a high-order interpolation solution becomes equal to an image filtering. The effect of this filter is somehow the same as the one of the histogram binning. Indeed, if some entries of the histograms are originally null, that is, if an intensity is not represented in the image, then filtering this image offers a greater opportunity to make this intensity appear and therefore smooths the MI function.

The curve  $\text{MI}_{\text{final}}$  in Fig. 2 shows the results obtained with a simple  $5 \times 5$  Gaussian filter on both images  $\mathbf{I}$  and  $\mathbf{I}^*$  in the previous translation example. It illustrates the advantages of this approach that yields a more smooth and convex MI registration function.

To validate the proposed formulation, i.e., the 1-D translational example, where MI was previously evaluated with noisy images, has also been performed in the occlusion and illumination variation cases. The values of the SSD are also computed to justify the use of the MI function. As Fig. 2 shows, MI, in the case of noise, occlusions, and illumination variations, remains robust while the SSD is not. Indeed, in the case of the occlusion, the link between the intensities of the nonoccluded part and the reference is stronger than any link between the new elements and the reference. Therefore, the optimum is unchanged. Considering the illumination variations, despite the modifications of the intensities, we keep the link between the intensities of the left part of the current and reference images as well as the link in the right part. Therefore, MI still provides an accurate estimation of the alignment position.

#### IV. MUTUAL INFORMATION IN VISUAL SERVOING

The goal is now to build the control law that will bring the pose of the camera to maximize the MI function to satisfy the problem as it is defined in Section II. Since the proposed MI is robust to many appearance variations, we can assume that the maximum of the MI will be reached when the current pose of the camera reaches the desired pose.

##### A. Mutual Information-Based Control Law

To perform the optimization and reach the maximum, we have to study the variation of the MI depending on the velocity of the camera.

The problem of finding the camera pose maximizing the MI can be reformulated as iteratively finding the velocity that brings the MI derivatives to a null value. These derivatives are computed with respect to the camera velocity which brings us to a problem of regulation of the interaction matrix of the MI. Using the formalism of [22], the regulation of a task function  $e$  to zero is done using the following control law:

$$\mathbf{v} = -\lambda \widehat{\mathbf{L}}_e^+ \mathbf{e}^\top \quad (14)$$

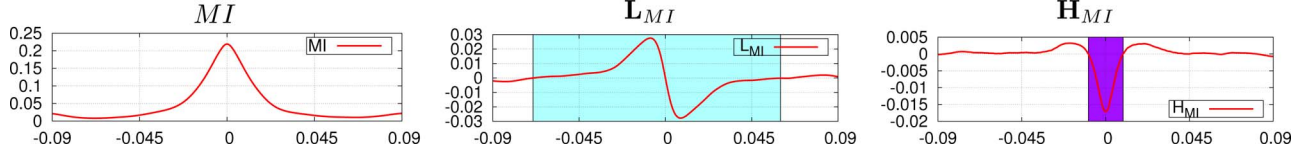


Fig. 4. MI and its derivatives in a nominal case with respect to the horizontal translation of the camera. The classical Newton's method has a thin convergence domain, that is, the concave domain represented in purple, while the proposed optimization method has a large convergence domain that is represented in sky blue.

where  $\lambda$  is a positive scalar factor used to tune the convergence rate, and  $\widehat{\mathbf{L}}_e^+$  is an estimation of the pseudoinverse of the interaction matrix associated with the task. In the classical geometric visual-servoing approaches, the task to regulate is the difference between the desired and current features. In our problem, we identify the task by  $\mathbf{L}_{\text{MI}}^\top$ , i.e., the interaction matrix of MI that has to be regulated to 0 since the gradient of MI is null at convergence. Since the task and the velocity have the same dimension, the pseudoinverse can be replaced by the inverse leading to

$$\mathbf{v} = -\lambda \mathbf{H}_{\text{MI}}^{-1} \mathbf{L}_{\text{MI}}^\top \quad (15)$$

where  $\mathbf{H}_{\text{MI}}$  is the interaction matrix of  $\mathbf{L}_{\text{MI}}$  that we call Hessian of MI. Given (10) and the chain rules simplifications detailed in [10], the expressions of the Gradient and Hessian are

$$\mathbf{L}_{\text{MI}} = \sum_{i,j} \frac{\partial p_{\Pi^*}}{\partial \mathbf{r}} \left( 1 + \log \left( \frac{p_{\Pi^*}}{p_{\text{I}}} \right) \right) \quad (16)$$

$$\begin{aligned} \mathbf{H}_{\text{MI}} &= \frac{\partial \mathbf{L}_{\text{MI}}}{\partial \mathbf{r}} \\ &= \sum_{i,j} \frac{\partial p_{\Pi^*}}{\partial \mathbf{r}}^\top \frac{\partial p_{\Pi^*}}{\partial \mathbf{r}} \left( \frac{1}{p_{\Pi^*}} - \frac{1}{p_{\text{I}}} \right) + \frac{\partial^2 p_{\Pi^*}}{\partial \mathbf{r}^2} \left( \frac{p_{\Pi^*}}{p_{\text{I}}} \right). \end{aligned} \quad (17)$$

For the purpose of clarity, we noted the interaction matrix of a variable  $x$  as  $\partial x / \partial \mathbf{r}$  where the correct notation should be  $\mathbf{L}_x$  [3]. It is often proposed in the literature [9], [10], [26], to approximate the Hessian matrix by neglecting the second-order derivatives. In our approach, we compute the full Hessian matrix using the second-order derivatives that are, in our point of view, required to obtain a precise estimation of the motion. More details are given in Appendix to highlight the problem caused by this classical approximation. Considering expressions (16) and (17), all the required variables are known apart from the joint probability derivatives. Using the joint probability expression given in (13) yields the following derivatives expressions:

$$\begin{aligned} \frac{\partial p_{\Pi^*}(i, j, \mathbf{r})}{\partial \mathbf{r}} &= \frac{1}{N_x} \sum_x \frac{\partial \phi}{\partial \mathbf{r}} (i - \bar{\mathbf{I}}(\mathbf{r}, \mathbf{x})) \phi(j - \bar{\mathbf{I}}^*(\mathbf{x})) \\ \frac{\partial^2 p_{\Pi^*}(i, j, \mathbf{r})}{\partial \mathbf{r}^2} &= \frac{1}{N_x} \sum_x \frac{\partial^2 \phi}{\partial \mathbf{r}^2} (i - \bar{\mathbf{I}}(\mathbf{r}, \mathbf{x})) \phi(j - \bar{\mathbf{I}}^*(\mathbf{x})). \end{aligned} \quad (18)$$

In order to compute the joint probability derivatives, the  $\phi$  function has to be two times differentiable. In our study, we consider that  $\phi$  is a B-spline function. To satisfy the necessary differentia-

bility condition,  $\phi$  is chosen as a third-order B-spline ( $\phi = B_3$ , see Fig. 3).

The interaction matrix of the function  $\phi$  can then be decomposed as

$$\frac{\partial \phi}{\partial \mathbf{r}} (i - \bar{\mathbf{I}}(\mathbf{r}, \mathbf{x})) = -\frac{\partial \phi}{\partial i} (i - \bar{\mathbf{I}}(\mathbf{r}, \mathbf{x})) \nabla \bar{\mathbf{I}} \mathbf{L}_x \quad (19)$$

where  $\nabla \bar{\mathbf{I}} = (\nabla \bar{\mathbf{I}}_{x_m}, \nabla \bar{\mathbf{I}}_{y_m}) = (p_x \nabla \bar{\mathbf{I}}_x, p_y \nabla \bar{\mathbf{I}}_y)$  is the gradient of the image  $\bar{\mathbf{I}}$  expressed in the metric space that are obtained using the classical image gradients and the camera intrinsic parameters  $(p_x, p_y)$  that is the ratio between the focal length and the size of a pixel.  $\mathbf{L}_x$  is the interaction matrix that links the displacement of a point in the image plan to the camera velocity. The interaction matrix is given by [3]

$$\mathbf{L}_x = \begin{bmatrix} -1/Z & 0 & x/Z & xy & -(1+x^2) & y \\ 0 & -1/Z & y/Z & 1+y^2 & -xy & -x \end{bmatrix}$$

where  $(x, y)$  are the coordinates of the point expressed in meters in the image plan, and  $Z$  is its depth relative to the camera. In this paper, we consider that the depth of the scene is unknown, and thus we simply set the depth of each point constant.

Using the same principle, the second-order derivative of the  $\phi$  function is given by

$$\begin{aligned} \frac{\partial^2 \phi}{\partial \mathbf{r}^2} (i - \bar{\mathbf{I}}(\mathbf{r}, \mathbf{x})) &= \frac{\partial^2 \phi}{\partial i^2} (i - \bar{\mathbf{I}}(\mathbf{r}, \mathbf{x})) (\nabla \bar{\mathbf{I}} \mathbf{L}_x)^\top (\nabla \bar{\mathbf{I}} \mathbf{L}_x) \\ &\quad - \frac{\partial \phi}{\partial i} (i - \bar{\mathbf{I}}(\mathbf{r}, \mathbf{x})) (\nabla \bar{\mathbf{I}}_x \mathbf{H}_x + \nabla \bar{\mathbf{I}}_y \mathbf{H}_y) \\ &\quad - \frac{\partial \phi}{\partial i} (i - \bar{\mathbf{I}}(\mathbf{r}, \mathbf{x})) \mathbf{L}_x^\top \nabla^2 \bar{\mathbf{I}} \mathbf{L}_x \end{aligned} \quad (20)$$

where  $\nabla^2 \bar{\mathbf{I}} \in \mathbf{R}^{2 \times 2}$  is the gradient of  $\nabla \bar{\mathbf{I}}$  in the metric space, and  $\mathbf{H}_x$  and  $\mathbf{H}_y$  are, respectively, the derivatives of the first and second line of the interaction matrix  $\mathbf{L}_x$  (see [13] for the computation of the two Hessian matrices).

The resulting MI, interaction matrix  $\mathbf{L}_{\text{MI}}$  and Hessian  $\mathbf{H}_{\text{MI}}$  are represented in Fig. 4 using the same 1-DOF approach as was used in Fig. 2.

## B. Optimization Approaches

Newton's method makes the assumption that the cost function to optimize is parabolic. Since MI is quasi-concave, this assumption is valid near the convergence where the function is concave. The definition of MI given in Section III-B yields a large concave domain and thus a large convergence domain. Nevertheless, a larger convergence domain can be obtained.

Indeed, if we focus on the MI and MI derivatives values reported in Fig. 4 for a simple 1-DOF example, we see that the concave domain is relatively small (the domain in purple

where the Hessian is negative) compared with the domain of convergence that a steepest gradient descent would have (sky blue domain). However, let us recall that the steepest gradient descent is not adapted for our 6-DOF problem since several of these DOF are highly correlated [5].

Usually, the optimization is improved with some line-search methods [28] or with the Levenberg–Marquardt approach [16], [20]. Nevertheless, those techniques require backtracking strategies that are not suitable to the visual-servoing problem since it is impossible to move the robot to test various positions.

The solution that we propose is to study the valley shape of the cost function at convergence. Knowing the shape, it is possible to modify the steepest gradient-descent direction to make it follow the valley. This optimization approach is commonly known as a preconditioning approach [2].

To characterize the valley shape at convergence, we simply estimate the Hessian matrix of the cost function computed at convergence. A good assumption is then simply to consider that the current image at convergence is similar to the desired image, and the Hessian matrix at convergence  $\mathbf{H}_{\text{MI}}^*$  is then given by (17) using  $\mathbf{I} = \mathbf{I}^*$ .

The resulting Hessian matrix is a negative matrix since it is computed at the maximum of the MI function and is ideal to adapt the direction of the gradient to make it follow the valley using

$$\mathbf{v} = -\lambda \mathbf{H}_{\text{MI}}^{*-1} \mathbf{L}_{\text{MI}}^{\top} \quad (21)$$

Since the Hessian matrix is computed using only the reference image, singularities are possible only when there is not enough information on the reference image (for instance, when there are no gradients (only null gradients) on the horizontal or vertical axes of the image). Moreover, the matrix  $\mathbf{H}_{\text{MI}}^*$  also gives an ideal norm to the velocity that brings it to a null value at convergence. We can observe that if we formulate the problem using the task function of [22], this approach is equal to approximating  $\widehat{\mathbf{L}}_{\mathbf{e}}$  by  $\mathbf{L}_{\mathbf{e}}^*$ , i.e., the interaction matrix of the task at the desired position, that is common in the geometric visual-servoing approaches [3].

Let us note that the time-consuming computation of the Hessian matrix is performed only once in this approach. Only the computation of the interaction matrix  $\mathbf{L}_{\text{MI}}$  is required at each iteration; thus, the control law computation is very fast.

## V. EXPERIMENTAL RESULTS

To validate the proposed approach, several experiments have been performed using a camera mounted on a 6-DOF gantry robot. Independently from the experiment, the computation time remains low. The control law is computed at video rate. A velocity is computed and sent to the robot every 20 ms for a  $320 \times 240$  input image using a 2.4-GHz computer.

### A. Positioning Tasks

A first set of experiments are realized to validate the robustness of the proposed approach using monomodal images in nominal conditions, as well as with occlusions and illumination variations to validate the robustness of MI. The camera is first moved to the desired pose  $\mathbf{r}^*$ , where the reference image is ac-

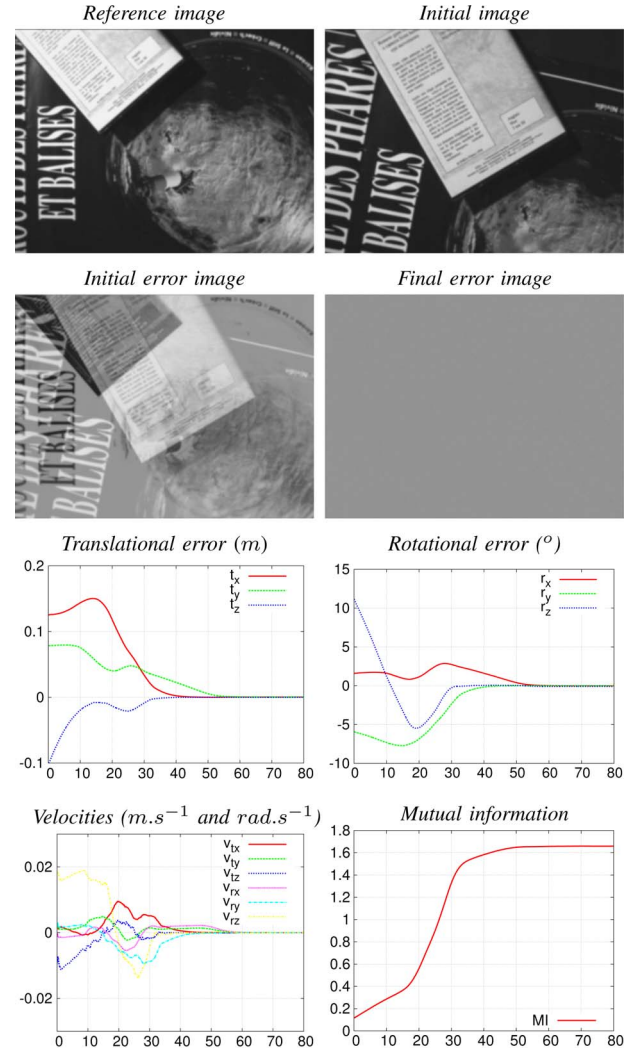


Fig. 5. Visual servoing using MI. The positioning error, velocities, and MI are represented with respect to the time in seconds.

quired. The camera pose is then set to an initial pose  $\mathbf{r}$ , which ensures that the reference image is partially represented in the current image.

During the task, the positioning error  $\Delta \mathbf{r}$ , which is the transformation between  $\mathbf{r}$  and  $\mathbf{r}^*$ , is computed to evaluate the behavior of the task. For a purpose of clarity, we will refer to the positioning error as  $\Delta \mathbf{r}_{\text{trans}}$ , i.e., the norm of the translation between the center of the camera at the current pose and at the desired pose in meters, and as  $\Delta \mathbf{r}_{\text{rot}}$ , i.e., the norm of rotation error in degrees.

1) *Nominal Conditions:* In this experiment, the illumination conditions remain constant during the realization of the positioning task. Fig. 5 shows the desired and initial images acquired by the camera, the initial and final error images, and the evolution of the positioning error using the cartesian coordinates for the translation part of  $\Delta \mathbf{r}$  and the error on the rotational part.

From an initial positioning error of  $\Delta \mathbf{r}_{\text{trans}} = 0.18$  m and  $\Delta \mathbf{r}_{\text{rot}} = 12.2^\circ$ , the camera is smoothly converging to the desired pose to reach a final positioning error of  $\Delta \mathbf{r}_{\text{trans}} = 3 \times 10^{-4}$  m and  $\Delta \mathbf{r}_{\text{rot}} = 0.06^\circ$ . Considering that the distance

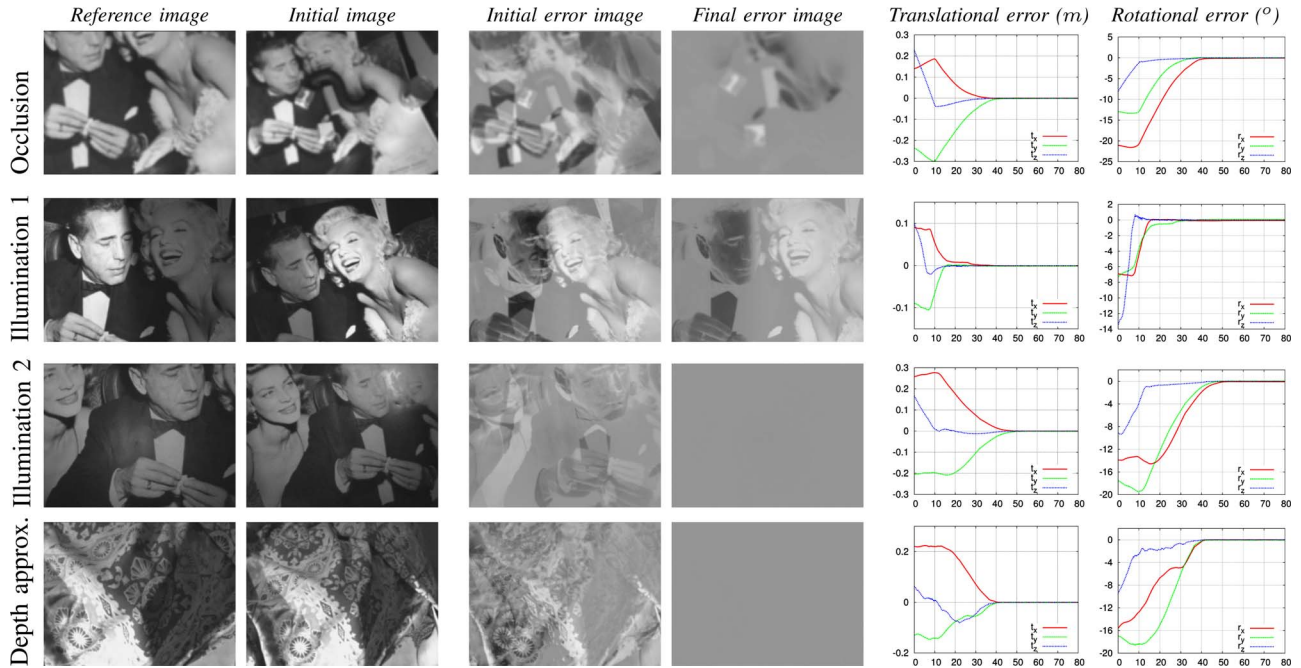


Fig. 6. Visual servoing using MI. The robustness of the visual servoing task with respect to occlusions, illumination variations, and depth approximation is verified by the evolution of the positioning error over time (in seconds) that converges to zero and by the appearance of the final error image.

from the camera to the scene is about 1 m, the proposed visual-servoing task proves to be very accurate compared with the accuracy that feature-based approaches would provide.

A second observation concerns the required degree of overlap between the current and desired images. This experiment has been chosen to illustrate the possible convergence domain that can be reached using the proposed method. In this experiment, only 50% of the reference image pixels are present in the initial image. Using this scene, we reach the limit of the convergence possibilities of the proposed visual-servoing task. A lower percentage would cause the task to fail, while a larger percentage yields to a success in most of the cases (avoiding the cases with large rotations and translations around the focal axis). A thorough analysis of the convergence domain is given in the next section.

2) *Robustness With Respect to Occlusions and Illumination Variations:* In these experiments, the appearance of the scene is modified during the positioning task using occlusions or illumination variations. Two experiments are illustrated to show the robustness of the proposed approach. In the first one (see Illumination 1 in Fig. 6), the illumination conditions are changed by moving the light sources in the environment, while in the second one (see Illumination 2), a light source is mounted on the camera, which leads to important moving specularities on the acquired images. As we can see in Fig. 6, despite the large initial positioning error and the change in appearance, the visual-servoing task converges and the positioning error decreases with respect to the time to become almost null. As expected, the proposed MI-based visual-servoing scheme is naturally robust to large perturbations and remains very accurate. Indeed, in the illumination variation experiments, the initial positioning error is  $\Delta \mathbf{r}_{\text{trans}} = 0.17$  m and  $\Delta \mathbf{r}_{\text{rot}} = 16.3^\circ$  in the first one and  $\Delta \mathbf{r}_{\text{trans}} = 0.37$  m



Fig. 7. External view of the scene (draperies) considered in the depth approximation experiment. The depth variation exceeds 30 cm.

and  $\Delta \mathbf{r}_{\text{rot}} = 24.5^\circ$  in the second. The visual-servoing task reaches a final positioning error of  $\Delta \mathbf{r}_{\text{trans}} = 9 \times 10^{-4}$  m and  $\Delta \mathbf{r}_{\text{rot}} = 0.08^\circ$ . In the occlusion experiment, the initial error is  $\Delta \mathbf{r}_{\text{trans}} = 0.35$  m and  $\Delta \mathbf{r}_{\text{rot}} = 25.2^\circ$ ; after the visual-servoing task, it is  $\Delta \mathbf{r}_{\text{trans}} = 1 \times 10^{-3}$  m and  $\Delta \mathbf{r}_{\text{rot}} = 0.1^\circ$ . The whole experiment using the first illumination variations is presented in the attached video.

3) *Robustness With Respect to Depth Approximation:* In the previous experiments, the desired image was always depicting a fronto-parallel scene. Therefore, the effect of the scene's depth approximation presented in Section IV was limited. In this section, an experiment is performed to illustrate the robustness of the proposed approach with respect to this approximation in a more extreme case. The considered scene is no longer planar, with a depth variation exceeding 30 cm (see the external view in Fig. 7), while the distance between the camera and the scene is about 1 m.

Despite the error that is introduced in the computation of the Gradient and Hessian matrices of the MI, the control law keeps converging to the desired position. Indeed, this approximation

only causes a small bias in the computation of the gradient, and the experiment shows that this bias is negligible. The fourth row of Fig. 6 shows the initial and desired images with the evolution of the positioning error. With an initial positioning error of  $\Delta \mathbf{r}_{\text{trans}} = 0.27$  m and  $\Delta \mathbf{r}_{\text{rot}} = 25.1^\circ$ , the final pose is very close to the desired pose with an accuracy equal to the one obtained in nominal conditions.

### B. Multimodal Image-Based Navigation

In the definition of the MI that we proposed, a linear dependence between the intensities of the desired and current images is not required. MI is thus able to align two images even if they are acquired from different modalities as soon as they share enough information. In this experiment, we show the robustness of MI with respect to multimodal alignment by servoing the camera on an aerial scene using a map reference image.

Here, we consider a different task: following a visual path. The goal is now to reproduce a trajectory using a sequence of previously learned reference images and one visual-servoing task per reference image. The challenge in this experiment is that the scenes during the learning and the navigation steps are from different modalities. Indeed, the sequence of reference image is learnt while the camera is moving over a map (provided by the Institut géographique national (IGN) géoportail), and the multiple visual-servoing tasks are performed over a satellite image (at the same scale).

Several navigation tasks have been performed with success (see also the results in [7]). Fig. 8 illustrates one of these navigation tasks. We can see that the current and reference images are correctly aligned with the displacement of the camera, despite their large differences of appearance (that would cause every feature-based or photometric-based technique to fail). Thus, the resulting camera trajectory of the navigation task properly re-plays the learned trajectory. The navigation experiment is presented in the attached video.

## VI. EMPIRICAL CONVERGENCE ANALYSIS

When considering IBVS with redundant features, only local stability can be considered, and this is also the case for our MI-based visual-servoing scheme. Nevertheless, it is always possible to evaluate the convergence area of this method from an empirical point of view. This section evaluates the performances of the proposed visual-servoing approach on a large set of simulated experiments (to consider simulation allows us to perform exhaustive tests with hundreds of positioning tasks).

### A. Convergence Domain and Performance Metrics

A set of initial poses have been chosen to best evaluate the robustness of the task with respect to the six DOF of the robot.

The initial poses are set so that the center of the camera is placed on a regular 3-D grid centered on the desired pose, and its direction is defined so that the initial and desired images overlap (this implies large variations around the  $r_x$  and  $r_y$  axes). In this case, each initial pose ensures that the current image shares some information with the desired image. We also consider a

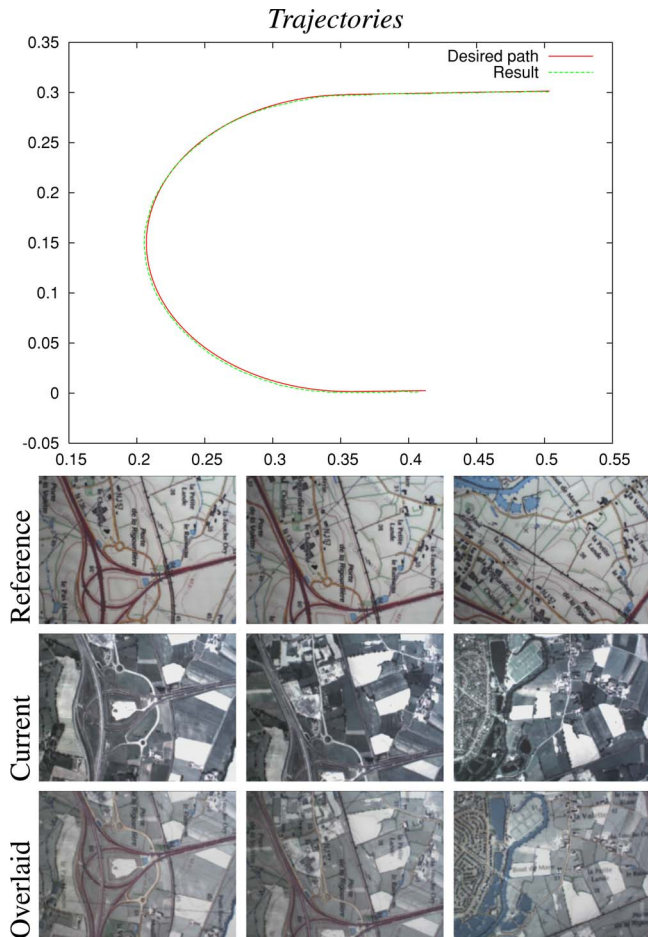


Fig. 8. Multimodal MI-based visual servoing in a navigation task. An image path is learned on the map scene and the visual-servoing task performs the navigation on the satellite scene. The reference and resulting trajectories are very close. The correct alignment is also visible in the image with the reference and current images overlaid.

high degree of rotation variation around the camera  $z$ -axis  $r_z$  since these rotations are usually difficult to handle in visual servoing.

Fig. 9 shows the convergence results obtained on a 3-D grid of  $21 \times 21 \times 21$  initial poses varying from  $-2$  to  $2$  m on the translations along the  $x$ - and  $y$ -axis (producing rotations from  $-60^\circ$  to  $60^\circ$  around the same axis) and from  $-1$  to  $1$  m in translation along the camera  $z$ -axis with an initial rotation  $r_z$  of  $0^\circ$  and  $20^\circ$ . The convergence domain is considerable. We can notice on the slices along the three planes that define the grid that the convergence domain is convex and that its hull has a spherical shape approximately centered on the desired pose with a radius of about 1 m.

Although the previous experiments described the convergence area, it is also of interest to analyze the camera trajectory during the positioning task. A set of four quantitative metrics have been measured on this set of experiments to perform this evaluation [11]. The first is the convergence ratio that gives the proportion of converging visual-servoing tasks on the whole set of experiments. The second and third are the average distance covered by the center of the camera and the average integral

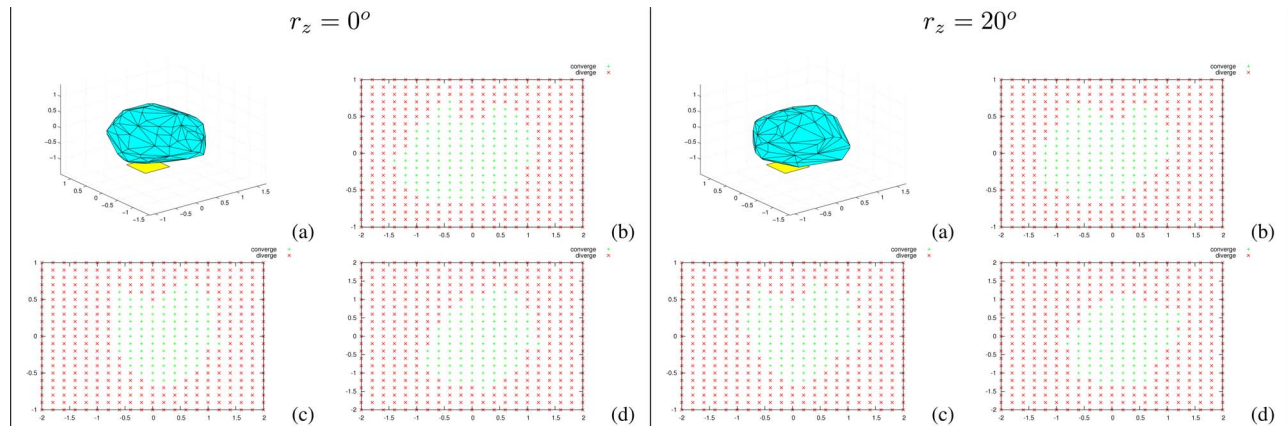


Fig. 9. Domain of attraction in the 3-D space with respect to the rotation around the focal axis. The blue volume in (a) represents the convex hull of the initial poses that converged. The yellow plan represents the target (a). Slices of the domain of attraction through the x(b), y(c), and z(d) planes are represented with the poses that converged in green and that diverged in red.

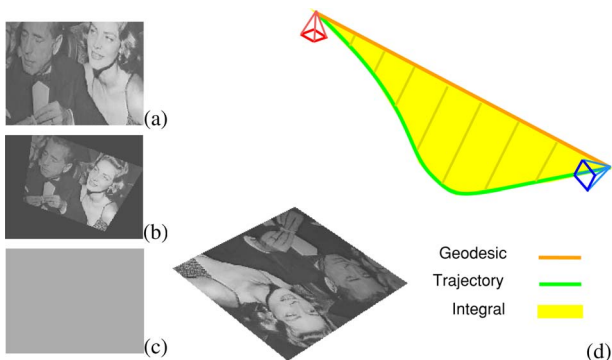


Fig. 10. Performance metrics on one task. (a) Reference image. (b) Image at the initial pose. (c) Final image error. (d) Resulting trajectory (green) of the camera from the initial pose (blue) to the desired pose (red).

between the camera center and the geodesic (we consider only successful experiments). Both measures are illustrated in Fig. 10 on the resulting trajectory of the camera in one of the experiments with an initial positioning error of  $\Delta \mathbf{r}_{\text{trans}} = 1$  m and  $\Delta \mathbf{r}_{\text{rot}} = 49.0^\circ$ . Finally, the last metric is the final positioning error, that is, the transformation between the final pose of the camera and its desired pose.

The results that have been obtained are represented in Table I with respect to the initial rotational error around the focal axis  $r_{z_0}$ . The proposed visual-servoing task has a large convergence domain and good performance measures. The greater the initial rotational error, the less important is the convergence rate. Indeed, if it is too large, then the initial amount of shared information between the images is too small, and the control law reaches a local optimum. The final positioning error has not been reported in the table since it remains constant with a final translation error of  $3 \times 10^{-5}$  mm and  $0.003^\circ$  in every converging experiment.

### B. Comparison With Existing Solutions

Let us now compare our approach with other visual-servoing schemes in one typical visual-servoing task with and with-

TABLE I  
PERFORMANCE OF THE PROPOSED VISUAL SERVOING TASK ON THE SET OF INITIAL POSES REPRESENTED IN FIG. 9

| $r_{z_0}$                 | $0^\circ$ | $10^\circ$ | $20^\circ$ | $30^\circ$ | $40^\circ$ | $50^\circ$ |
|---------------------------|-----------|------------|------------|------------|------------|------------|
| Convergence (%)           | 26.7      | 27.6       | 25.2       | 21.0       | 12.4       | 1.8        |
| Distance (m)              | 1.06      | 1.08       | 1.11       | 1.16       | 1.27       | 1.65       |
| Integral ( $\text{m}^2$ ) | 0.11      | 0.12       | 0.12       | 0.13       | 0.14       | 0.25       |

out illumination variations. Two visual-servoing approaches, which are adapted to the current problem, have been considered. The first one is the photometric-based visual servoing [5]. The second one is a classical feature-based visual servoing where the features are points extracted and matched using the scale-invariant feature transform (SIFT) algorithm [14].

The obtained results are summarized in Fig. 11 where SSD refers to the photometric approach. Without illumination variations, the proposed approach has a similar behavior to the photometric one. In terms of trajectory, the direct approaches (i.e., MI and photometric) are further from the geodesic than the SIFT approach. Indeed, the optimization in the SIFT-based approach is performed on a quasi-quadratic function, while the cost function in the direct approaches are much more nonlinear yielding the presented trajectories. Considering the final positioning error, the direct approaches are more accurate than the feature-based approaches. The accuracy of the direct approaches comes from the fact that no intermediate measure is considered. In the SIFT approach, the coordinates of the extracted points are intermediate measures that cause measurement errors and limit the accuracy of the positioning task. Furthermore, in direct methods, all the information contained in the image is considered, and this redundancy allows for greater improvement regarding the positioning accuracy. A good alternative is then to use the feature-based approach and switch the control law at convergence to finally use the MI-based visual-servoing approach to take advantage of both a trajectory near the geodesic and a very accurate final pose. This approach, which we call hybrid approach, has been implemented and gives indeed the most adapted behavior with both advantages of trajectory and accuracy.

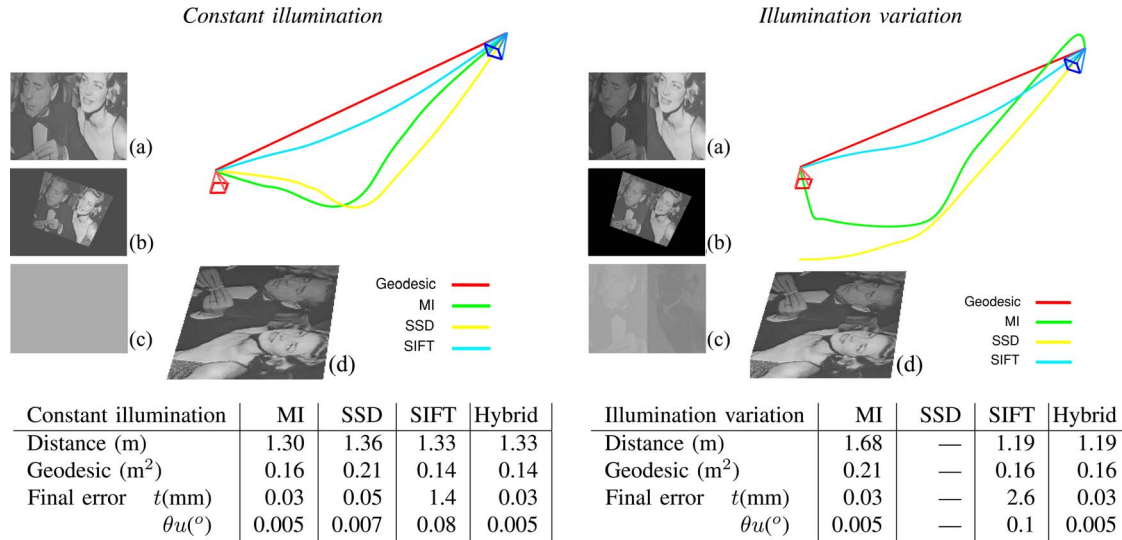


Fig. 11. Comparison between our MI-based VS, the photometric VS (SSD), an SIFT-based VS, and an hybrid VS with and without illumination variations. (a) Reference image. (b) Image acquired at the initial pose. (c) Final error image. (d) Trajectories in the 3-D space. The tables show the corresponding values of the performance metrics.

To evaluate the performance of the approaches with respect to illumination variations, the following variation has been applied to the scene: From the acquisition of the desired image to the visual-servoing tasks, the left part of the scene has been illuminated and the right part is put in the shadow. The consequence of such a modification in the intensities causes the SSD function to have a minimum at the wrong camera pose, and thus, the photometric approach diverges. The illumination variation has also a slight effect on the matching step of the SIFT approach; the visual-servoing task is then converging but has a larger final positioning error. As for the MI-based approach, the trajectory of the camera is slightly affected, but the final positioning error remains very small.

## VII. CONCLUSION

In this paper, we presented a new visual-servoing approach based on MI. This new control law does not use any feature; therefore, it also does not require any extraction, matching, or tracking step that are usually the bottleneck of classical approaches.

The goal is to bring the image acquired from the camera to be the more similar to a reference image. Thus, only the reference image has to be known to reach the desired position. Moreover, since the similarity measure is the MI, the new control law is naturally robust to partial occlusions and illumination variations. Another advantage, which comes from the fact that it is a featureless approach, is that there is no measurement errors due to feature extraction, and thus, the positioning task is very accurate.

As is well known in the medical field, MI is also robust to multimodal alignment. Some new visual-servoing applications are, therefore, possible, including, for instance, aerial drone navigation. Although our experiments were limited to map and aerial images, other modalities can be easily considered such as infrared images.

## APPENDIX

### WHY THE HESSIAN MATRIX MUST NOT BE APPROXIMATED

It is common to find the Hessian matrix of MI given in (17) approximated by the following expression [10], [26]:

$$\mathbf{H}_{\text{MI}} \simeq \sum_{i,j} \frac{\partial p_{\Pi^*}}{\partial \mathbf{r}}^\top \frac{\partial p_{\Pi^*}}{\partial \mathbf{p}} \left( \frac{1}{p_{\Pi^*}} - \frac{1}{p_{\mathbf{I}^*}} \right) \quad (22)$$

where the second-order derivative of the joint probability has been neglected. The approximation is inspired from the one that is made in the Gauss–Newton method for a least-squared problem. It assumes that the second-order derivative is null at the convergence.

Considering the expression of the marginal probability  $p_{\mathbf{I}^*}(j) = \sum_i p_{\Pi^*}(i, j)$ , it is clear that  $p_{\mathbf{I}^*}(j) > p_{\Pi^*}(i, j)$ ; therefore,  $1/p_{\Pi^*}(i, j) - 1/p_{\mathbf{I}^*}(j) > 0$ . Since  $\frac{\partial p_{\Pi^*}}{\partial \mathbf{r}}^\top \frac{\partial p_{\Pi^*}}{\partial \mathbf{r}}$  is a positive matrix, the final Hessian matrix given by (22) is positive. Since the optimum of MI is a maximum, the Hessian matrix at convergence is supposed to be negative by definition. The common approximation of (22) is thus not suited for the optimization of MI.

## REFERENCES

- [1] A. H. A. Hafez, S. Achar, and C. V. Jawahar, "Visual servoing based on gaussian mixture models," in *Proc. IEEE Int. Conf. Robot. Autom.*, Pasadena, CA, May 2008, pp. 3225–3230.
- [2] J. F. Bonnans, J. Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal, *Numerical Optimization—Theoretical and Practical Aspects*. Berlin, Germany: Springer-Verlag, 2006.
- [3] F. Chaumette and S. Hutchinson, "Visual servo control. Part I: Basic approaches," *IEEE Robot. Autom. Mag.*, vol. 13, no. 4, pp. 82–90, Dec. 2006.
- [4] G. Chesi and K. Hashimoto, Eds., *Visual Servoing via Advanced Numerical Methods*. New York: Springer-Verlag, 2010.
- [5] C. Collewet and E. Marchand, "Photometric visual servoing," *IEEE Trans. Robot.*, 2011, to be published.
- [6] A. Dame and E. Marchand, "Entropy based visual servoing," in *Proc. IEEE Int. Conf. Robot. Autom.*, Kobe, Japan, May 2009, pp. 707–713.

- [7] A. Dame and E. Marchand, "Improving mutual information based visual servoing," in *Proc. IEEE Int. Conf. Robot. Autom.*, Anchorage, AK, May 2010, pp. 5531–5536.
- [8] K. Deguchi, "A direct interpretation of dynamic images with camera and object motions for vision guided robot control," *Int. J. Comput. Vis.*, vol. 37, no. 1, pp. 7–20, Jun. 2000.
- [9] N. Dowson and R. Bowden, "Mutual information for Lucas-Kanade tracking (milk): An inverse compositional formulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 180–185, Jan. 2008.
- [10] N. D. H. Dowson and R. Bowden, "A unifying framework for mutual information methods for use in non-linear optimisation," in *Proc. Eur. Conf. Comput. Vis.*, Jun. 2006, vol. 1, pp. 365–378.
- [11] N. R. Gans, S. Hutchinson, and P. I. Corke, "Performance tests for visual servo control systems, with application to partitioned approaches to visual servo control," *Int. J. Robot. Res.*, vol. 22, no. 10–11, pp. 955–984, 2003.
- [12] V. Kallem, M. Dewan, J. P. Swensen, G. D. Hager, and N. J. Cowan, "Kernel-based visual servoing," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, San Diego, CA, Oct. 2007, pp. 1975–1980.
- [13] J. T. Lapresté and Y. Mezour, "A Hessian approach to visual servoing," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sendai, Japan, Sep. 2004, pp. 998–1003.
- [14] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
- [15] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.
- [16] F. Maes, D. Vandermeulen, and P. Suetens, "Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information," *Med. Image Anal.*, vol. 3, no. 4, pp. 373–386, 1999.
- [17] E. Malis, "Improving vision-based control using efficient second-order minimization techniques," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 2, New Orleans, LA, Apr 2004, pp. 1843–1848.
- [18] E. Marchand and F. Chaumette, "Feature tracking for visual servoing purposes," *Robot. Auton. Syst.*, vol. 52, no. 1, pp. 53–70, Jun. 2005 (Special Issue on "Advances in Robot Vision," D. Kragic, H. Christensen, Eds.).
- [19] S. K. Nayar, S. A. Nene, and H. Murase, "Subspace methods for robot vision," *IEEE Trans. Robot.*, vol. 12, no. 5, pp. 750–758, Oct. 1996.
- [20] G. Panin and A. Knoll, "Mutual information-based 3D object tracking," *Int. J. Comput. Vis.*, vol. 78, no. 1, pp. 107–118, 2008.
- [21] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual information matching and interpolation artefacts," in *SPIE Medical Imaging*, vol. 3661, K. M. Hanson, Ed. Bellingham, WA: SPIE, 1999, pp. 56–65.
- [22] C. Samson, B. Espiau, and M. Le Borgne, *Robot Control: The Task Function Approach*. London, U.K.: Oxford Univ. Press, 1991.
- [23] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, Dec. 1979.
- [24] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.
- [25] H. A. Sturges, "The choice of a class interval," *Amer. Statist. Assoc.*, vol. 21, pp. 65–66, 1926.
- [26] P. Thévenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Trans. Image Process.*, vol. 9, no. 12, pp. 2083–2099, Dec. 2000.
- [27] P. Viola and W. Wells, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, 1997.
- [28] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," in *Proc. IEEE Int. Conf. Comput. Vis.*, Washington, DC, 1995, p. 16.



servoing.



**Amaury Dame** received the Eng. degree from the Institut National des Sciences Appliquées de Rennes, Rennes, France, in 2007, the STI Master's degree in signal and image processing from the Université de Rennes 1, in 2007, and the Ph.D. degree in computer science from the University of Rennes in 2010, where he did research with the INRIA Lagadic Team under the supervision of E. Marchand.

He is currently with the Lagadic Team, CNRS, INRIA Rennes—Bretagne Atlantique, IRISA. His research interests include computer vision and visual

**Eric Marchand** received the Ph.D. and "Habilitation à Diriger des Recherches" degrees in computer science from the University of Rennes, Rennes, France, in 1996 and 2004, respectively.

He is currently a Professor of computer science with the Université de Rennes 1, and a member of the INRIA Lagadic team. He spent one year as a Post-doctoral Associate the AI Laboratory, Department of Computer Science, Yale University, New Haven, CT. He was an INRIA Research Scientist ("Chargé de recherche") at IRISA-INRIA Rennes from 1997 to

2009. His research interests include robotics, perception strategies, visual servoing, real-time object tracking, and augmented reality.

Dr. Marchand has been an Associate Editor for the IEEE TRANSACTIONS ON ROBOTICS since 2010.