Entropy-Based Visual Servoing

Amaury Dame, Eric Marchand

Abstract— In this work we propose a new way to achieve visual servoing using directly the information (as defined by Shannon) of the image. A metric derived from information theory, mutual information, is considered. Mutual information is widely used in multi-modal image registration (medical applications) since it is insensitive to changes in the lighting condition and to a wide class of non-linear image transformation. In this paper mutual-information is used as a new visual feature for visual servoing and allows us to build a new control law to control the 6 dof of the robot. Among various advantages, this approach does not require any matching nor tracking step, is robust to large illumination variation and allows to consider, within the same task, different image modalities. Experiments that demonstrate these advantages conclude the paper.

I. Introduction

A. Motivations

Visual servoing consists in using the information provided by a vision sensor to control the movements of a dynamic system [2]. This approach requires to extract information (usually geometric features) from the image in order to design the control law. Robust extraction and real-time spatio-temporal tracking of these visual cues [9] is a non trivial task and also one of the bottlenecks of the expansion of visual servoing.

In [4], it has been shown that no other information than the image intensity (the pure image signal) can 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 variation.

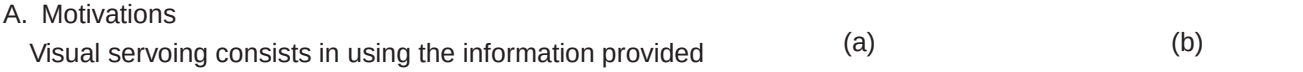
In this paper, we propose a new approach that no longer relies on geometrical features [2] nor on pixels intensity [4] but use directly the information (entropy) contained in the image signal. More precisely we will consider mutual information [16]. Being closer from the signal, we will show that this new approach

- is robust to very important light variations (see Figure 1a),
- is robust to important occlusions,
- is able to consider different image modalities (see Figure 1b).

B. Overview and related works

Classically, to achieve a visual servoing task, a set of visual features has to be selected from the image allowing to control the desired degrees of freedom (dof). A control law has also to be designed so that these visual features reach a desired value s^*, leading to a correct realization of the task. The control principle is thus to regulate to zero the error vector $s - s^*$. To build this control law, the knowledge of the interaction matrix L_s, that links the time variation of s to the camera instantaneous velocity, is usually required [2]. Nevertheless, the key point of this approach is the choice of the visual features s. With a vision sensor providing 2D measurements $x(r_k)$ (where r_k is the camera pose at time k), potential visual features s are numerous, since 2D data (coordinates of feature points in the image, moments, ...) as well as 3D data provided by a localization algorithm exploiting the extracted 2D features can be considered. If the choice of s is important, it is always designed from visual measurements $x(r_k)$. A robust extraction, matching (between $x(r_k)$ and the desired measurements $x^* = x(r^*)$) and real-time spatio-temporal tracking (between $x(r_{k-1})$ and $x(r_k)$) have proved to be difficult, as testified by the abundant literature on the subject. These tracking and matching processes are even more difficult when acquisition configuration is modified during the execution of the task or if two different sensors or acquisition modalities are considered.

Recently different approaches have been proposed to get over these issues by considering no longer geometric features but the image itself or a function of the image. Considering the whole image as a feature avoids the tracking and matching process. Following this way, various approaches have been presented. [5], [10] consider the full image but in order to reduce the dimensionality of image data they



(a) (b)

Fig. 1. The visual servoing scheme considering mutual information as visual feature is able to handle very important lighting variation (a). Furthermore, different modalities can be considered for image acquisition (b). First row is desired image while second row shows the image acquired by the robot. In (a), the positioning task was correctly achieved despite an important modification of the lighting condition between the learning step and the execution of the task (see section IV-A for details). In (b) the learning step was done using a map while the servoing task was carried out on the corresponding aerial images. It is nevertheless possible to carry out a trajectory tracking task (see section IV-B).

Amaury Dame is with CNRS, IRISA, Lagadic team, Rennes, France. Eric Marchand is with INRIA Rennes - Bretagne Atlantique, IRISA, Lagadic team, Rennes, France. This work is supported by DGA under contribution to student grant. firstname.name@irisa.fr.

consider an eigenspace decomposition of the image. The control is then performed directly in the eigenspace which requires the off-line computation of this eigenspace (using a principal component analysis) and then, for each new frame, the projection of the image on this subspace. To cope with these issues a way to compute the interaction matrix related to the luminance under temporal luminance constancy case has been proposed in [4]. In that case, the error to be regulated is nothing but the sum of squared differences (SSD) between the current and the desired images $\| \mathbf{I} - \mathbf{I}^* \|$. Such approach is nevertheless quite sensitive to illumination variations (although using a more complex illumination model in some particular cases is possible [3]). [7] also considers the pixels intensity. This approach is based on the use of kernel methods that lead to a high decoupled control law. However, only the translations and the rotation around the optical axis are considered. Another approach that does not require tracking nor matching has been proposed in [1]. It models collectively feature points extracted from the image as a mixture of Gaussian and tries to minimize the distance function between the Gaussian mixture at current and desired positions. Simulation results show that this approach is able to control the 3 dof of the robot. However, note that an image processing step is still required to extract the current feature points.

As stated considering image intensity is quite sensitive to modification of the environment. To solve problems due to illumination changes or multi-modal servo, information contained in the images is considered and no more directly the luminance. The feature is the mutual information defined by Shannon in [12]. The mutual information (built from the entropy) of two random variables is a quantity that measures the mutual dependence of the two variables. Considering two images, the higher the mutual information is, the better is the alignment between the two images. Considering information contained in the image and not the image itself allows to be independent from perturbation or from the image modality. Such approach has been widely used for multi-modal medical image registration [16] [8] and more recently in tracking [6].

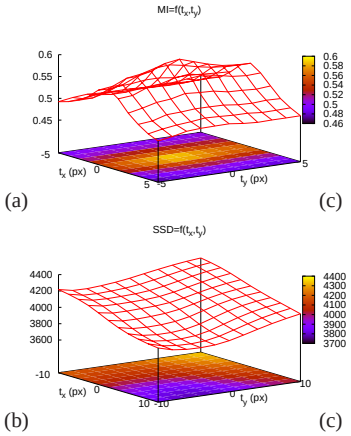
The remainder of this paper is organized as follows. In section II a background on information theory is presented and mutual information is formulated related to images. The resulting control law is presented in section III. Finally experiments on a 6 dof robot are presented in section IV.

II. Information Theory

The feature considered in previous works was the SSD which only deals with quasi identical images. To extend capabilities of the servoing task, information between images is considered. In this section entropy, joint entropy and mutual information are generally defined to end with the use of mutual information on images.

A. Mutual information

1) Entropy: To understand mutual information, a brief definition of entropy of a random variable is required. The entropy $H(X)$ of a random variable X (image, signal...) is mostly used in signal compression: it defines the theoretical number of bits needed to encode a random variable. If x are



(a) (c)

(b) (c)

Fig. 2. Illumination changes. Value of the mutual information (c) and SSD (d) by translating image (a) in the image space $(t_x; t_y)$ and comparing it with image (b) from the same position with illumination changes. SSD has a minimum in $(t_x; t_y) = (-2; -2)$ while mutual information has a correct maximum in $(t_x; t_y) = (0; 0)$.

the possible values of X and $p_X(x) = P(X = x)$, then the entropy $H(X)$ is given by:

$$H(X) = -\sum_x p_X(x) \log_2 (p_X(x)) . \quad (1)$$

By definition $0 \log_2(0) = 0$. For legibility issues $\log$ will be used as $\log_2$. The more values x are equally probable the more entropy $H(X)$ is bigger.

2) Joint entropy: Following the same idea joint entropy $H(X, Y)$ of two random variables X and Y can be defined as:

$$H(X, Y) = -\sum_{x,y} p_{X,Y}(x, y) \log (p_{X,Y}(x, y)) \quad (2)$$

where x and y are respectively the possible values of X and Y, $p_{X,Y}(x, y) = P(X = x \cap Y = y)$ is the joint probability of the values x and y. Typically the joint entropy defines the theoretical number of bits needed to encode a joint system of two random variables. At first sight finding the minimum of this entropy can be seen as an alignment method. But the dependencies on entropies of X and Y is a problem. In fact by adding a variable to another it is impossible to make the global entropy decrease, so $\min(H(X, Y)) = \max(H(X), H(Y))$.

For example considering a signal X, if a signal $Y = X$ is added to the system, the system will keep the same entropy $H(X, Y) = H(Y)$. Y does not add variability to the system. Now if we add a constant signal Y, the system keep the same entropy for the same reason. But in the second situation the two signals cannot be considered as aligned.

3) Mutual information: The definition of mutual information solve the above mentioned problem [12]. Mutual information of two random variables X and Y is given by the following equation:

$$MI(X, Y) = H(X) + H(Y) - H(X, Y). \quad (3)$$

Using equations (1) and (2) it yields to:

$$MI(X, Y) = \sum_{x,y} p_{xy}(x, y) \log \left(\frac{p_{xy}(x, y)}{p_x(x) p_y(y)} \right) \quad (4)$$

As shown in this equation, the dependencies on the entropies are suppressed by the difference between random variable's entropies and joint entropy. Mutual information is then the quantity of information shared between two random variables. If mutual information is maximized, then the two signals are aligned. The advantage of this function compare to SSD is that no linear relation is needed between the two signals [15]. To illustrate possibilities of alignment, mutual information has been computed applying a translation to images of different illumination conditions (Figure 2). A maximum at zero translation (the alignment position) is shown using mutual information whereas the SSD leads to an incorrect result.

B. Mutual information on images

In previous section mutual information has been defined for every kind of random variables. Now our interest is to use it to compare two images.

If $\mathbf{I} = \mathbf{I}(\mathbf{r})$ represents the image at the current pose $\mathbf{r}$ of the camera and if $\mathbf{I}^* = \mathbf{I}(\mathbf{r}^*)$ is the image at the desired pose $\mathbf{r}^*$ ($\mathbf{r}$ and $\mathbf{r}^*$ both elements of $\mathbb{R}^3 \times SO(3)$), using previous equations, mutual information of two images $\mathbf{I}$ and $\mathbf{I}^*$ is given by:

$$MI\left(\mathbf{I}(\mathbf{r}), \mathbf{I}^*\right) = \sum_{i,j} p_{ij}(i,j,\mathbf{r}) \log\left(\frac{p_{ij}(i,j,\mathbf{r})}{p_i(i,\mathbf{r}) p_j(j)}\right) \quad (5)$$

where i and j are respectively the pixel luminances allowed in the images $\mathbf{I}$ and $\mathbf{I}^*$. Typically the number of gray levels $N_{c_\mathbf{I}}$ and $N_{c_{\mathbf{I}^*}}$ of the images $\mathbf{I}$ and $\mathbf{I}^*$ are 256 $\left((i,j) \in [0;255]^2 \subset \mathbb{Z}^2\right)$. $p_i(i,\mathbf{r})$ and $p_j(j)$ are respectively the probability of the luminance i and j in the images $\mathbf{I}$ and $\mathbf{I}^*$. Knowing that $\mathbf{r}^*$ is constant, for clarity issue, in the remainder of this paper $p_j(j)$ and $p_{ij}(i,j,\mathbf{r})$ will respectively denote $p_j(j,\mathbf{r}^*)$ and $p_{ij}(i,j,\mathbf{r},\mathbf{r}^*)$. The probabilities can simply be computed as a normalized histogram of the images:

$$p_i(i,\mathbf{r}) = \frac{1}{N_\mathbf{x}} \sum_\mathbf{x} \delta_0\left(i - \mathbf{I}(\mathbf{x},\mathbf{r})\right) \quad (6)$$

$$p_j(j) = \frac{1}{N_\mathbf{x}} \sum_\mathbf{x} \delta_0\left(j - \mathbf{I}^*(\mathbf{x})\right) \quad (7)$$

where $N_\mathbf{x}$ is the number of pixels in the region of interest of the image, $\delta(x)$ is a Kronecker's function: $\delta(x) = 1$ for $x = 0$ else $\delta(x) = 0$.

$p_{ij}(i,j,\mathbf{r})$ is the joint probability of the two luminances i and j computed using a normalization of the joint histogram of the images:

$$p_{ij}(i,j,\mathbf{r}) = \frac{1}{N_\mathbf{x}} \sum_\mathbf{x} h(i,j,\mathbf{r}) \quad (8)$$

$$= \frac{1}{N_\mathbf{x}} \sum_\mathbf{x} \delta_0\left(i - \mathbf{I}(\mathbf{x},\mathbf{r})\right) \delta_0\left(j - \mathbf{I}^*(\mathbf{x})\right)$$

where $h(i,j,\mathbf{r})$ is the joint intensity histogram of the two images.

Considering every gray levels of the images, mutual information has been computed using translation around the zero position. As shown in Figure 3, the maximum is very sharp giving an accurate result. However, the shape of the cost function outside the maximum is quite planar, causing possible artefacts in case of noise.

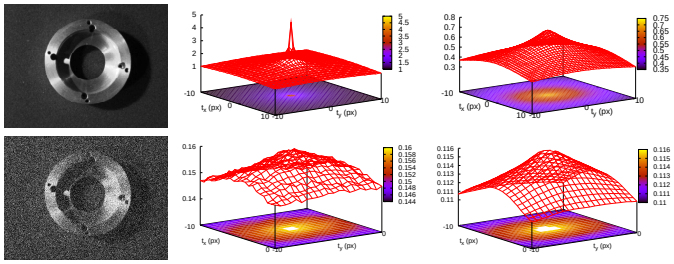


Fig. 3. Influence of the bin-size of the histogram Nc. Value of the mutual information between the image of the first column and its translation in the image space (t_x;t_y). Second column: original mutual information (Nc = 256), third column: mutual information with a bin-size of histogram Nc = 8. First row: without noise, second row: adding Gaussian noise to the image and its translation.

To overcome this problem, the in-Parzen windowing formulation of MI is used [13]. The luminances of the two initial images are divided to fit in the desired space of values $[0; Nc-1] \subset \mathbb{R}$ where Nc is the new bin-size of the histogram h. Let $\overline{\mathbf{I}}$ and $\overline{\mathbf{I}^*}$ represent the new images:

$$\overline{\mathbf{I}}(\mathbf{x}) = \mathbf{I}(\mathbf{x}) \frac{N_c}{N_{c_\mathbf{I}}} \qquad \overline{\mathbf{I}^*}(\mathbf{x}) = \mathbf{I}^*(\mathbf{x}) \frac{N_c}{N_{c_{\mathbf{I}^*}}}. \quad (9)$$

The only difference concerning the equation of mutual information (Eq. 5) is that the summation is no more on 256 but on Nc values. The principal changes occur in the computation of the marginal and joint probability. To keep most information despite quantifying, a B-spline function is used:

$$p_i(i,\mathbf{r}) = \frac{1}{N_\mathbf{x}} \sum_\mathbf{x} \phi\left[i - \overline{\mathbf{I}}(\mathbf{x},\mathbf{r})\right] \quad (10)$$

$$p_j(j) = \frac{1}{N_\mathbf{x}} \sum_\mathbf{x} \phi\left[j - \overline{\mathbf{I}^*}(\mathbf{x})\right] \quad (11)$$

$$p_{ij}(i,j,\mathbf{r}) = \frac{1}{N_\mathbf{x}} \sum_\mathbf{x} \phi\left[i - \overline{\mathbf{I}}(\mathbf{x},\mathbf{r})\right] \phi\left[j - \overline{\mathbf{I}^*}(\mathbf{x})\right] \quad (12)$$

A detailed description of B-spline functions is given by Unser *et al.* in [14] but interesting properties of B-spline are recalled here: the integral of the function being 1, the result does not have to be renormalized and the computation of the derivatives is easily obtained. To keep a low computational cost, in the following experiments a B-spline of order 2 has been selected:

$$\phi(t) = \begin{cases} t+1 & \text{if } t \in [-1,0] \\ -t+1 & \text{if } t \in [0,1] \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

In Figure 3 a computation of mutual information is presented using a histogram's bin-size of $Nc = 256$ and one with $Nc = 8$ on two identical images applying a translation. It shows that the maximum is wider with a smaller Nc adding noise robustness to the optimisation problem.

III. Visual Servoing based on Mutual Information

Having a robust alignment function, now the goal is to use it to reach the desired pose $\mathbf{r}^*$ with the camera i.e. to maximize information mutual to the current and desired frame. In this section to respect convention of minimization the opposite of mutual information is used:

$$\mathbf{r}^* = \min_{\mathbf{r}} \left(-MI(\mathbf{I}(\mathbf{r}), \mathbf{I}^*)\right). \tag{14}$$

This alignment problem brings us to an optimization problem. Having the camera at current position $\mathbf{r}_t$, the gradient of the cost function is used to find $\mathbf{v} = (\boldsymbol{\upsilon}, \boldsymbol{\omega})$ the velocity vector in the Cartesian space applied to the camera to reach position $\mathbf{r}_{t+1}$ corresponding to a higher mutual information. Each minimization step can be written as follows:

$$\mathbf{r}_{t+1} = \mathbf{r}_t \oplus \mathbf{v} \tag{15}$$

where "$\oplus$" defines the operator that applies a velocity to a pose. For the same reasons as in [4] the optimization method chosen in the following experiments is a Levenberg-Marquardt like approach which allows to smoothly pass from Gauss-Newton to Steepest Descent method, depending on how far is the minimum:

$$\mathbf{v} = -\lambda(\mathbf{H} + \mu \text{diag}\mathbf{H})^{-1}\mathbf{G}^\top \tag{16}$$

where $\mathbf{G} \in \mathbb{R}^{1\times 6}$ and $\mathbf{H} \in \mathbb{R}^{6\times 6}$ are respectively the Gradient and the Hessian of the cost function. As explained in [6], from (5) the Gradient can be computed this way:

$$\begin{aligned} \mathbf{G} &= -\frac{\partial MI(\mathbf{I}(\mathbf{r}), \mathbf{I}^*)}{\partial \mathbf{r}} \\ &= -\sum_{i,j} \frac{\partial p_{ij}}{\partial \mathbf{r}} \left(1 + \log\left(\frac{p_{ij}}{p_i}\right)\right). \end{aligned} \tag{17}$$

The derivative of the joint probability $\partial p_{ij}/\partial \mathbf{r} \in \mathbb{R}^{1\times 6}$ is computed using:

$$\frac{\partial p_{ij}}{\partial \mathbf{r}} = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \frac{\partial \phi}{\partial \mathbf{r}}(i - \overline{\mathbf{I}}(\mathbf{x}, \mathbf{r}))\phi(j - \overline{\mathbf{I}^*}(\mathbf{x})) \tag{18}$$

where $\partial\phi/\partial\mathbf{r} \in \mathbb{R}^{1\times 6}$ is:

$$\frac{\partial \phi(i - \overline{\mathbf{I}}(\mathbf{x}, \mathbf{r}))}{\partial \mathbf{r}} = -\frac{\partial \phi(i - \overline{\mathbf{I}}(\mathbf{x}, \mathbf{r}))}{\partial i} \frac{\partial \overline{\mathbf{I}}(\mathbf{x}, \mathbf{r})}{\partial \mathbf{r}}$$

if the system is considered Lambertian, then the variation of the image luminance from the camera position can be decomposed as follows:

$$\begin{aligned} \frac{\partial \phi(i - \overline{\mathbf{I}}(\mathbf{x}, \mathbf{r}))}{\partial \mathbf{r}} &= -\frac{\partial \phi(i - \overline{\mathbf{I}}(\mathbf{x}, \mathbf{r}))}{\partial i} \frac{\partial \overline{\mathbf{I}}(\mathbf{x}, \mathbf{r})}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{r}} \\ &= -\frac{\partial \phi(i - \overline{\mathbf{I}}(\mathbf{x}, \mathbf{r}))}{\partial i} \nabla \overline{\mathbf{I}}\, \mathbf{L}_{\mathbf{x}} \end{aligned} \tag{19}$$

where $\nabla\overline{\mathbf{I}}$ is nothing but the image gradient $\left(\frac{\partial \overline{\mathbf{I}}(\mathbf{x},\mathbf{r})}{\partial x}, \frac{\partial \overline{\mathbf{I}}(\mathbf{x},\mathbf{r})}{\partial y}\right)$ and $\mathbf{L}_{\mathbf{x}}$ is the interaction matrix at the point $\mathbf{x} = (x, y)$. Using a perspective projection it leads to:

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

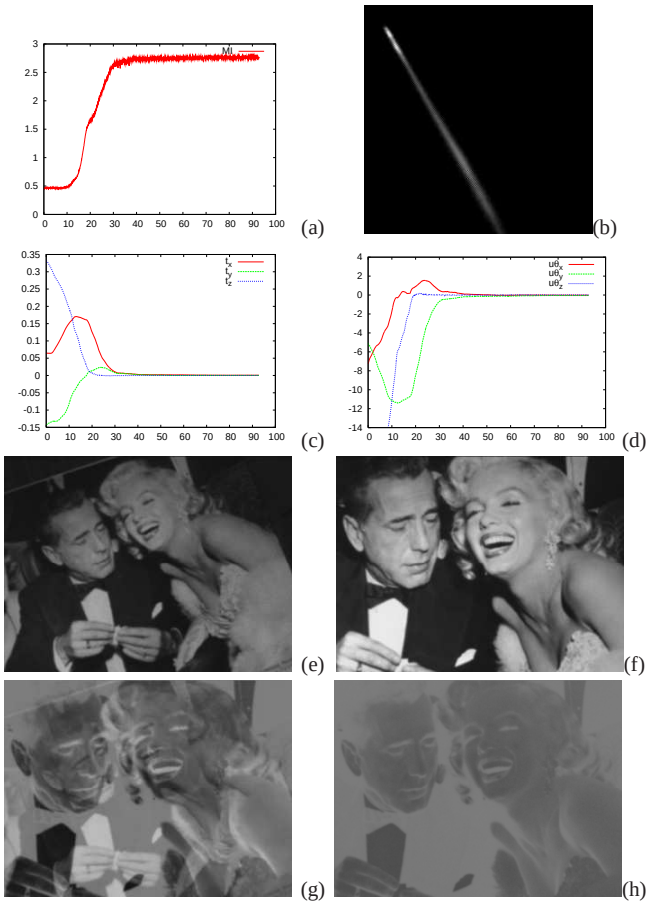


Fig. 4. First experiment: global illumination changes. (a) Mutual information, (c) translation part of $\mathbf{r}$ (in meter) and (d) rotational part of $\mathbf{r}$ (°) with x axis in seconds. (b) Final joint histogram, (e) initial image, (f) desired image, (g) initial images difference and (h) final images difference $\mathbf{I}^* - \mathbf{I}$.

where Z is the depth of the point relative to the camera frame and x and y are the coordinates of the point in the image frame depending on the camera intrinsic parameters. Given the equation of $\mathbf{G}$, the Hessian $\mathbf{H}$ is given by:

$$\begin{aligned} \mathbf{H} &= \frac{\partial \mathbf{G}}{\partial \mathbf{r}} \\ &= -\sum_{i,j} \frac{\partial p_{ij}}{\partial \mathbf{r}}^\top \frac{\partial p_{ij}}{\partial \mathbf{r}} \left(\frac{1}{p_{ij}} - \frac{1}{p_i}\right) + \frac{\partial^2 p_{ij}}{\partial \mathbf{r}^2} \left(\frac{p_{ij}}{p_i}\right) \\ &\simeq -\sum_{i,j} \frac{\partial p_{ij}}{\partial \mathbf{r}}^\top \frac{\partial p_{ij}}{\partial \mathbf{r}} \left(\frac{1}{p_{ij}} - \frac{1}{p_i}\right). \end{aligned} \tag{20}$$

The last term of the second equation is quasi null near the desired position and is very expensive to compute. Since it is usual in visual servoing to compute the interaction matrix at the desired position [2], this term is neglected in the following experiments without affecting the convergence.

IV. Experimental results

All the experiments reported in this paper have been obtained using a camera mounted on the end-effector of a six dof gantry robot. Computation time is 22ms for each 320×240 frames using a 2.6 Ghz PC.

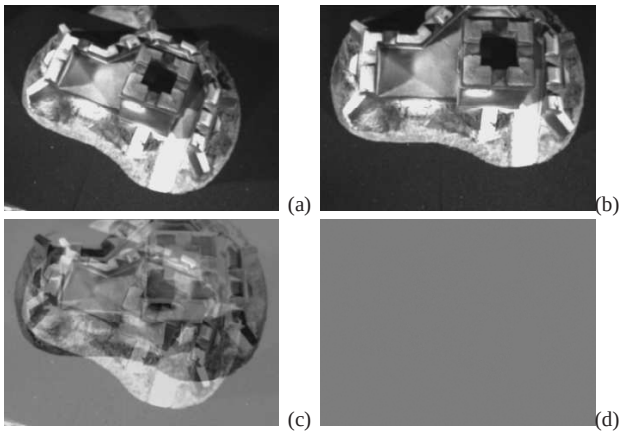


Fig. 5. Depth approximation. (a) Initial image, (b) desired image, (c) initial images difference and (d) final images difference $\mathbf{I}^* - \mathbf{I}$.

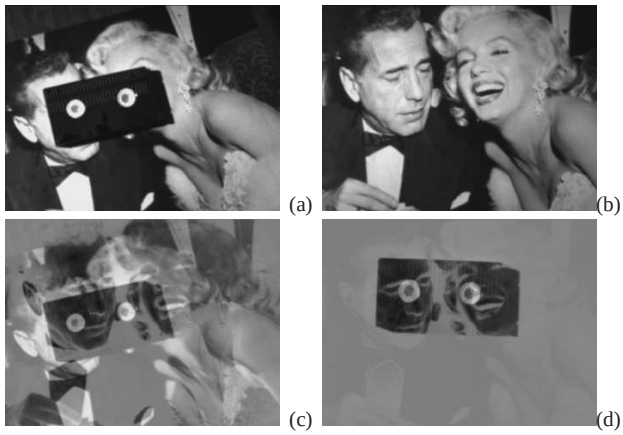


Fig. 6. Occlusions robustness. (a) Initial image, (b) desired image, (c) initial images difference and (d) final images difference $\mathbf{I}^* - \mathbf{I}$.

A. Visual servoing positioning experiments

A set of experiments shows the behavior of our approach for positioning task. In each cases, the manipulator is first moved to the desired pose $\mathbf{r}^*$ and the corresponding picture $\mathbf{I}^*$ is acquired. The manipulator is then moved to its initial pose $\mathbf{r}$. The control signals computed using equation (16) are sent to the robot controller until convergence. To validate the quality of the results, the transformation $\Delta\mathbf{r}$ between $\mathbf{r}$ and $\mathbf{r}^*$ is computed and analyzed.

1) General analysis: We will first consider the behavior of the algorithm using a planar object so that the object and the image planes are parallel at the desired pose. The initial error pose is $\Delta\mathbf{r}_{init} = (15\text{cm}, -15\text{cm}, 30\text{cm}, -6°, -6°, 18°)$. The global illumination of the entire scene has been modified during the realization of the task. Figure 4 pictures the results of the first experiment. Here an approximation of the depth of the plane at the desired pose is known, then interaction matrix are computed using a constant depth $Z =$ 70cm at each point. Results are quite accurate: Figure 4 (c) and (d) show the pose error between the desired and final positions during the servoing task. The final pose error $\Delta\mathbf{r}$ is $(0.1\text{mm}, -0.1\text{mm}, -0.1\text{mm}, 0.01°, -0.01°, -0.01°)$. The final images difference $\mathbf{I}(\mathbf{r}_{\mathbf{final}}) - \mathbf{I}^*$ is not null since the global illumination has been modified. However the alignment can be shown in the image representing the joint histogram between the images $\mathbf{I}(\mathbf{r}_{final})$ and $\mathbf{I}^*$: along the axes the luminances of the two images are plotted, from left to right for final image and from top to bottom for the desired image. The feature space is constructed by counting the number of times a combination of grey values occurs. For each pair of corresponding points (x, y), with x a point in the image $\mathbf{I}$ at final pose $\mathbf{r}_{final}$ and y a point in the desired image $\mathbf{I}^*$, the entry $(\mathbf{I}(x, \mathbf{r}_{final}), \mathbf{I}^*(y))$ in the feature space is increased. Using this representation (See Figure 4 (b)) a quasi linear relation between $\mathbf{I}(\mathbf{r}_{final})$ and $\mathbf{I}^*$ is visible, depicting an alignment between the two images with a decreased illumination in $\mathbf{I}(\mathbf{r}_{final})$.

Let us note that at the beginning of the experiment the bin-size of the histogram h is set to $Nc = 8$, increasing the domain of convergence, and the parameter μ of the Levenberg-Marquardt method is set to $\mu = 0.1$, favouring a steepest descent approach. Using this set of parameters during all the experiment leads to approximate results: the first issue encountered is that the current pose reaches a valley of the cost function which the steepest descent does not deal with (as in [4]). Consequently an error remains on the couples of movements (t_x, θ_y) and (t_y, θ_x). The second problem is that considering a small Nc yields to a less precise minimum. Parameters μ and Nc are then modified during minimization. A polynomial filter is considered to detect the local minimum of the cost function. In such case, parameters μ and Nc are smoothly updated, increasing Nc and decreasing μ (leading to a Gauss-Newton minimization process).

2) Robustness wrt depth approximation: To support the use of a constant depth in the computation of the interaction matrix an experiment on non-planar object has been released. The initial pose error is $(8\text{cm}, -8\text{cm}, 8\text{cm}, -4°, -5°, 18°)$. The behaviour of the positioning task remains almost the same than the previous experiment. Result still shows a low positioning error of $(0.2\text{mm}, -0.1\text{mm}, -0.1\text{mm}, 0.01°, -0.01°, -0.01°)$ (See Figure 5).

3) Robustness wrt occlusion: The next experiment (scene similar to the first one) deals with a large partial occlusion. An object (video tape) is added to the scene after the learning step. Despite the introduction of the object in the scene, the manipulator is still moving toward the desired position, as in previous experiments. Finally the translation error is about $(0.1\text{mm}, -0.1\text{mm}, -0.1\text{mm})$ and the rotational error of $(0.01°, -0.01°, -0.01°)$ showing robustness to occlusions as expected (See Figure 6).

4) Robustness wrt large illumination changes: The goal of the last positioning experiment illustrates the robustness to large and non global illumination changes that SSD based approaches can not deal with. Light configuration is widely modified during the realization of the task leading to non-uniform lighting variation. The configuration allows to light independently various parts of the scene. The different illumination conditions in the image at the desired and initial poses is shown in Figure 7: at the desired pose the left part of the scene is illuminated and in the initial pose it is the right part. Then at the alignment position, the right part of the current image is a brighter version of

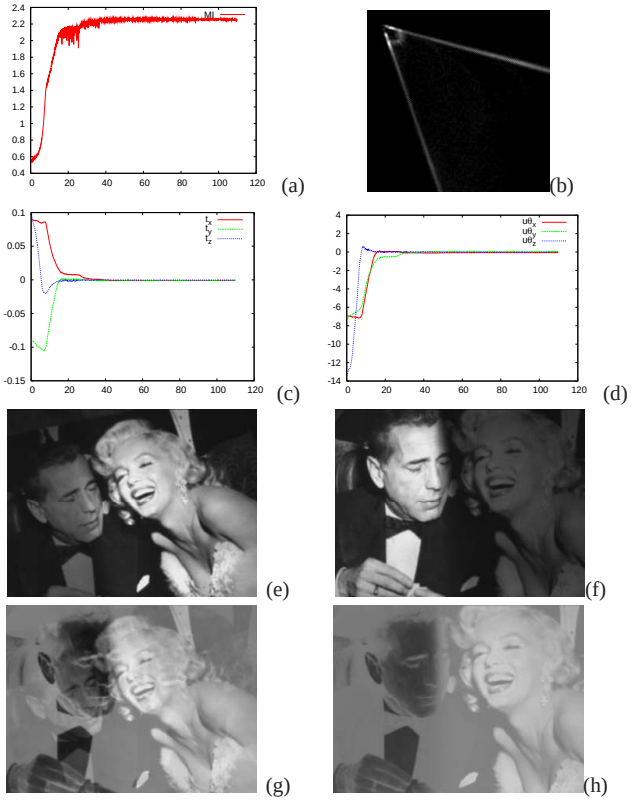


Fig. 7. Robustness to illumination changes. (a) Mutual information, (c) translation part of $\mathbf{r}$ (meter) and (d) rotational part of $\mathbf{r}$ (°) with x axis in seconds. (b) Final joint histogram, (e) initial image, (f) desired image, (g) initial images difference and (h) final images difference $\mathbf{I}^* - \mathbf{I}$.

the desired image, and the left part of the current image is a darker version of the desired image. As in the first experiment the images difference is not null. However the alignment can be seen in the final joint histogram image: two lines are visible corresponding to two quasi linear relations. One for each part of the image. Then the use of mutual information that deals with non linear relation between the images totally makes sense in such conditions while conventional SSD minimization would fail. Finally the manipulator reaches the desired pose with a final pose error of $(0.4\text{mm}, -0.4\text{mm}, -0.1\text{mm}, -0.05°, 0.06°, 0.01°)$ that are rather accurate results.

Another experiment has been tested using light changes during the servoing task. A presentation of these experiments is given in the video accompanying this paper.

B. Multimodal image-based navigation using image memory

In the introduction of this paper we suggested that the proposed approach is able to consider multi-modal images. To illustrate this property, we will consider an image-based navigation task that uses image memory. Following [11], we consider that the navigation task is defined in the sensor space by a database of images acquired during a learning step. This defines an image path which provides enough information to control the robotic system. In this system the desired image $\mathbf{I}^*$ used in (14) will vary with time. The cost function to minimize is then

$$\mathbf{r}^* = \min_{\mathbf{r}} \left(-MI(\mathbf{I}(\mathbf{r}), \mathbf{I}^*(t))\right). \tag{21}$$

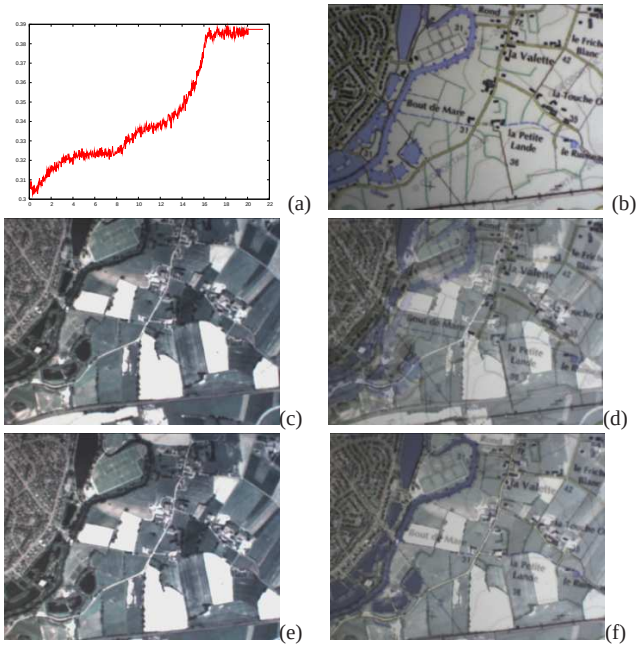


Fig. 8. A multi-modal servoing task. (a) Mutual information wrt time (seconds), (b) desired image, (c) initial image, (d) initial image overlaid on the desired image, (e) final image, (f) final image overlaid on the desired image.

The next desired image $\mathbf{I}^*(t)$ is taken in the database when the gradient of the mutual information between $\mathbf{I}(\mathbf{r})$ and $\mathbf{I}^*(t)$ is bellow a given threshold.

To illustrate the ability of our approach to consider multi-modal images, the learning step was used on a 1:25000 map while the navigation was used on aerial image. These map and aerial image have been acquired using the *IGN (National Institute of Geography) geoportail* (http://www.geoportail.fr) which is a tool similar to google earth. Map and aerial images have the same scale.

During the servoing task the aerial images were considered. The current and desired images are then very different (see Figure 8). Considering these extreme conditions, most of the intensity-based or feature-based matching techniques between current and desired images would fail. Nevertheless considering mutual-information, experiment shows very good results. The behavior of the visual servoing considering multimodal capabilities is shown on Figure 8. Figure 8a shows the desired image (a map) while Figures 8c and 8e show initial and final image acquired by the camera. Figures 8d and 8f show the desired image overlaid on the current one. On Figure 8f, one can see the registration between desired and final image has been precisely achieved. Figure 8a shows the value of the mutual information that increase during the servoing task.

Figure 9 shows five sets of images with the desired images extracted from the database and modified over time (top), current image acquired by the camera (middle), and the desired image overlaid on the current one to show the quality of the registration, and thus of the trajectory tracking process (bottom). Figure 10 shows both the learnt trajectory and the trajectory obtained during the realization of the navigation task. A presentation of these experiments is also available in

Fig. 9. Multi-modal visual servoing in a navigation task. First row: desired images (acquired during the learning step) ; second row: current image ; third row : desired image overlaid on the current one.

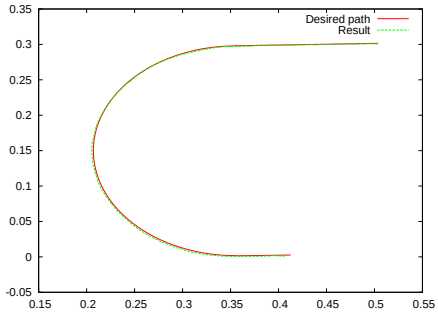


Fig. 10. Reference path along with actual camera displacement. x axis represents t_x translation and y axis t_y translation in meters.

the video accompanying this paper.

A typical application for this techniques would be aerial drones navigation. Although we consider here a map and aerial images, other modalities can be easily considered such as satellite images (visible or infrared layers), etc.

V. Conclusion

In this paper we presented a new metric for visual servoing. This metric, the mutual information between two images, is derived from the information theory (as defined by Shannon). A new control law, which does not required any matching nor tracking step, based on mutual information has been proposed. An explicit formulation of the interaction related to the mutual information is given.

Based on the information contained in the image, this approach is then insensitive to most of image perturbations and a variety of non-linear transformations for which most of the intensity-based or feature-based matching or tracking techniques between current and desired image would fail. In particular it is very robust to large illumination variation or occlusion. Furthermore, it features good behaviour concerning multi-modal visual servoing with possible applications in navigation.

References

[1] A.H. Abdul Hafez, S. Achar, and C.V. Jawahar. Visual servoing based on gaussian mixture models. In *IEEE Int. Conf. on Robotics and Automation, ICRA'08*, pages 3225–3230, Pasadena, California, May 2008.

[2] F. Chaumette and S. Hutchinson. Visual servoing and visual tracking. In B. Siciliano and O. Khatib, editors, *Handbook of Robotics*, chapter 24, pages 563–583. Springer, 2008.

[3] C. Collewet and E. Marchand. Modeling complex luminance variations for target tracking. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition, CVPR'08*, Anchorage, Alaska, June 2008.

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

[5] K. Deguchi. A direct interpretation of dynamic images with camera and object motions for vision guided robot control. *Int. Journal of Computer Vision*, 37(1):7–20, June 2000.

[6] N.D.H. Dowson and R. Bowden. A unifying framework for mutual information methods for use in non-linear optimisation. In *European Conf. on Computer Vision*, pages 365–378, 2006.

[7] V. Kallem, M. Dewan, J.P. Swensen, G.D. Hager, and N.J. Cowan. Kernel-based visual servoing. In *IEEE/RSJ Int. Conf. on Intelligent Robots and System, IROS'07*, pages 1975–1980, San Diego, USA, October 2007.

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

[9] E. Marchand and F. Chaumette. Feature tracking for visual servoing purposes. *Robotics and Autonomous Systems*, 52(1):53–70, June 2005. special issue on "Advances in Robot Vision", D. Kragic, H. Christensen (Eds.).

[10] S.K. Nayar, S.A. Nene, and H. Murase. Subspace methods for robot vision. *IEEE Trans. on Robotics*, 12(5):750 – 758, October 1996.

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

[12] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.

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

[14] Michael Unser, Akram Aldroubi, Murray Eden, and Life Fellow. B-spline signal processing: Part i-theory. *IEEE Trans. Signal Processing*, 41:821–833, 1993.

[15] P. Viola. *Alignment by Maximization of Mutual Information*. PhD thesis, MIT, 1995.

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