Proceedings of the 2001 IEEE
International Conference on Robotics & Automation
Seoul, Korea • May 21-26, 2001

Stacking Jacobians Properly in Stereo Visual Servoing

P. Martinet

LASMEA - GRAVIR
Blaise Pascal University of Clermont-Ferrand
63177 Aubière - Cedex, France
martinet@lasmea.univ-bpclermont.fr

E. Cervera

Robotic Intelligence Laboratory
Jaume-I University
12071 Castelló, Spain
ecervera@inf.uji.es

Abstract

Most visual servoing applications are concerned with geometrically modeled objects. In this paper, the problem of controlling a motion by visual servoing around an unknown object with a stereovision system is addressed. The main goal is to move the end-effector around the object in order to observe several viewpoints of the object for other tasks, e.g. inspection or grasping. The present work uses the well-known image-based visual servoing approach with a point, but the importance of the relationship between the end-effector and camera frames is clarified and emphasized. This relationship is needed for properly stacking the Jacobians or interaction matrices of each camera. A comparison with a visual servoing approach with a direct stacking of the Jacobians is presented. The centroid of a region, obtained by color segmentation, is used to move around the observed object. Experiments are developed on a PA-10 robot connected to a real time stereovision system, with two cameras mounted on the end-effector. Experimental results demonstrate the importance of a proper definition of the stacked Jacobians, to avoid undesired motions in the servoing task. Particularly, when turning around an unknown object, undesired motions on roll angle of the stereovision system can be avoided.

1 Introduction

Visual servoing applications have grown significantly since the last decade. Though in the first approaches, the scene observed by the camera was relatively simple, many works concerning unknown and complex objects have been developed recently.

Some methods require an initial learning step to obtain information characterizing the interaction between the sensor apparatus and the environment [3, 6]. In this case, it is necessary to get information from a predefined trajectory. To do so, the method proposed by Berry *et al.* [1] performs automatic motions around an unmodeled object in order to learn this interaction.

In this paper, the problem of moving around an object is addressed: no geometric model is needed and a stereovision system is used. Many works have been done in the field of visual servoing using a stereovision system [2, 5, 10]. Most of them use the stereovision system to recover the depth. Others use the epipolar constraint in order to execute the point to point matching process.

Recent developments in stereo-visual servoing have proved the theoretical soundness of the approach. Lamiroy *et al.* [7] present a solution to integrate the epipolar constraint directly in the control law. They rewrite the minimization problem under the optimization of the epipolar constraint, and show that, in the noiseless case and using rigid control points, both the classical and constrained approaches are identical.

Malis *et al.* [8] have formalized a *multi-cameras* visual servoing approach. They consider a system with N cameras which delivers a set $\mathbf{s} = (\mathbf{s}_1^T \mathbf{s}_2^T \cdots \mathbf{s}_N^T)^T$ of sensor signals ($dim(\mathbf{s}_i) = n_i$). Assuming that each sensor signal can control all the end-effector d.o.f m ($m \leq n_i$), they rewrite the global interaction relationship as:

$$\begin{aligned} \dot{\mathbf{s}} &= \begin{pmatrix} \mathbf{L}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{L}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{L}_N \end{pmatrix} \begin{pmatrix} {}^1\mathbf{M}_e \\ {}^2\mathbf{M}_e \\ \vdots \\ {}^N\mathbf{M}_e \end{pmatrix} {}^e\mathbf{v} \\ &= \mathbf{L}\mathbf{M}_e {}^e\mathbf{v} \end{aligned} \quad (1)$$

where $\mathbf{L}_i$ represents the interaction matrix (or Jacobian matrix) of the i^{th} camera, and ${}^i\mathbf{M}_e$ the transformation matrix between the velocity of the i^{th} camera and the robot end-effector velocity. They define a global task function $\mathbf{e} = \mathbf{C}\dot{\mathbf{s}} = \sum_{i=1}^{N} k_i \mathbf{e}_i$ as a weighted mean of the task function relative to each camera, and demonstrate some properties in convergence and stability.

In our work, this scheme is applied to a stereo rig composed of two cameras. We show that less d.o.f.

0-7803-6475-9/01/$10.00© 2001 IEEE

can be controlled, by defining an appropriate hybrid task.

The paper outline is as follows: first, the modeling aspect is developed; secondly, the *task-function* approach is applied to obtain the control law. Next, results obtained at video rate with our robotic platform vision system are shown. Finally, some conclusions and possible extensions are presented.

2 Modeling

The main goal of this work is the positioning of the end-effector with respect to a fixed object, and to perform motions around it. A stereo rig is rigidly attached to the end-effector. No model of the object is known a-priori, thus limiting the choice of features for tracking [1]. Nevertheless, the position of the cameras in the end-effector frame and its intrinsic parameters are roughly known, without any special calibration.

The proposed approach uses 2D features extracted from regions in the image, segmented by color. Such features can be the centroid of the region, its size, the aspect ratio, and the angle of its first axis of inertia.

In this first work, only a point feature (the centroid of the blob) is used. Its observation by the stereo pair mounted on the end-effector makes possible the 3D positioning task. At the same time, the 3 remaining d.o.f are used for a secondary task, e.g. moving the end-effector around the object while keeping the fixed relative position.

Our stereovision system is composed of two parallel cameras. Figure 1 illlustrates the case when both cameras observe a 3D point P.

Let us define $\mathcal{F}_e$ as the Cartesian frame attached to the end-effector, $\mathcal{F}_l$ as the frame attached to the left camera, and $\mathcal{F}_r$ as the frame attached to the right camera.

The feature vector is defined as $\mathbf{s} = (u_l, v_l, u_r, v_r)^T$ where $(u_l, v_l)^T$ and $(u_r, v_r)^T$ are the image coordinates of the point, observed by the left and right cameras respectively.

2.1 First control law: *real stereo*

Let ${}^e\mathbf{v}$ be the kinematic screw applied to the robot end-effector. According to the multi-cameras visual servoing formulated in equation (1), the relationship between the time derivative of the feature vector and the end-effector screw is

$$\dot{\mathbf{s}} = \begin{pmatrix} \mathbf{L}_l {}^l\mathbf{M}_e \\ \mathbf{L}_r {}^r\mathbf{M}_e \end{pmatrix} {}^e\mathbf{v} = \mathbf{L}_{st} {}^e\mathbf{v} \tag{2}$$

where

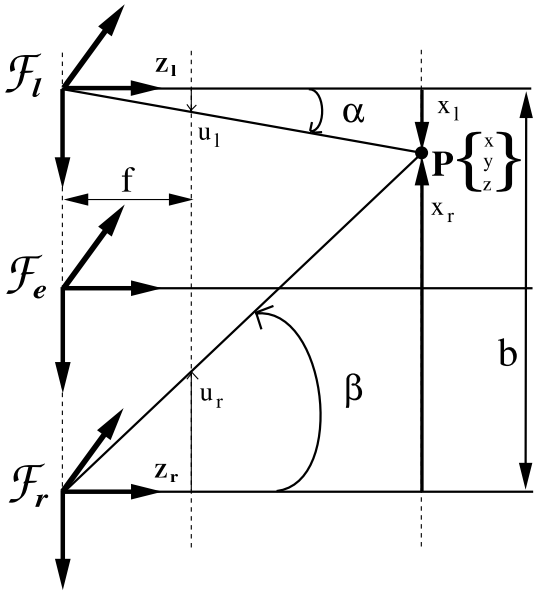

Figure 1: Stereovision: Case of a $3D$ Point.

- $\mathbf{L}_l$ and $\mathbf{L}_r$ are the interaction matrices relative to the left and right cameras respectively, defined by $(i = r \text{ or } l)$:

$$\begin{pmatrix} -\frac{F_u}{z} & 0 & \frac{u_i}{z} & \frac{u_i v_i}{F_v} & -F_u - \frac{u_i^2}{F_u} & \frac{v_i F_u}{F_v} \\ 0 & -\frac{F_v}{z} & \frac{v_i}{z} & F_v + \frac{v_i^2}{F_v} & -\frac{u_i v_i}{F_u} & -\frac{u_i F_v}{F_u} \end{pmatrix}$$

- ${}^l\mathbf{M}_e$ and ${}^r\mathbf{M}_e$ are the transformation matrices of the screw between the left and right camera frames and the end-effector frame. Given frames $\mathcal{F}_e$ and $\mathcal{F}_i$, the relationship between the screws is

$${}^i\mathbf{v} = {}^i\mathbf{M}_e {}^e\mathbf{v} \tag{3}$$

where the transformation matrix ${}^i\mathbf{M}_e$ is

$${}^i\mathbf{M}_e = \begin{pmatrix} {}^i\mathbf{R}_e & \left[{}^i\mathbf{t}_e\right]_\times {}^i\mathbf{R}_e \\ \mathbf{O}_3 & {}^i\mathbf{R}_e \end{pmatrix} \tag{4}$$

Though the resulting interaction matrix $\mathbf{L}_{st}$ is the same as that obtained by Maru *et al.* [10], our development is somewhat simpler and it is easier to generalize to other configurations of the cameras.

2.2 Second control law: *stacked-mono*

It is widely accepted in monocular visual servoing that the interaction matrix (the jacobian) of a set of points is constructed by stacking every interaction matrix of each single point.

One is tempted to apply this method directly to stereo vision, and thus, a simpler interaction matrix is obtained, if both matrices ${}^l\mathbf{M}_e$ and ${}^r\mathbf{M}_e$ are neglected.

In this case, the fusion of the sensor information is processed directly in the interaction matrix despite of the frame where they are defined. So, a more *classical* form is obtained:

$$\dot{\mathbf{s}} = \begin{pmatrix} \mathbf{L}_l \\ \mathbf{L}_r \end{pmatrix} {}^e\mathbf{v} = \mathbf{L}_{sm}\, {}^e\mathbf{v} \tag{5}$$

The interaction matrix is similar to that obtained by stacking the matrices of several points, hence the name of the control law.

2.3 Theoretical comparison

It can be shown that the null space of the stereo interaction matrix $\mathbf{L}_{st}$ is always spanned by the three vectors

$$\begin{pmatrix} 0 \\ z \\ -y \\ 1 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} -z \\ 0 \\ x \\ 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} y \\ -x \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{6}$$

which are in fact the same for the null space of the interaction matrix associated to a 3D point $\mathbf{p}_e = (x, y, z)^T$ in the robot end-effector frame

$$\begin{aligned} \mathbf{L}_{3D} &= \begin{pmatrix} -\mathbf{I}_3 & [\mathbf{p}]_\times \end{pmatrix} \\ &= \begin{pmatrix} -1 & 0 & 0 & 0 & -z & y \\ 0 & -1 & 0 & z & 0 & -x \\ 0 & 0 & -1 & -y & x & 0 \end{pmatrix} \end{aligned} \tag{7}$$

The interaction matrix $\mathbf{L}_{st}$ can be rewritten as:

$$\mathbf{L}_{st} = \mathbf{L}^{st}_{3D} \mathbf{L}_{3D} \tag{8}$$

where the matrix $\mathbf{L}^{st}_{3D}$ is defined as:

$$\mathbf{L}^{st}_{3D} = \begin{pmatrix} \frac{\partial \mathbf{s}_l}{\partial^l \mathbf{p}} {}^l\mathbf{R}_e \\ \frac{\partial \mathbf{s}_r}{\partial^r \mathbf{p}} {}^r\mathbf{R}_e \end{pmatrix} \tag{9}$$

thus it is composed of the partial derivatives of the image points with respect to the velocity of the 3D point. The matrix is full rank, i.e., its null space is empty, since there is no motion of the 3D point which leaves both images unaffected. Indeed, only the motion along the projection ray leaves an image constant, but, since the cameras are not coincident, their projection rays are obviously different. Consequently, the null space of matrix $\mathbf{L}_{st}$ is the same as that of $\mathbf{L}_{3D}$.

In our experiments, only one point is used, thus 3 d.o.f. of the end-effector remain free for moving around the object.

In the second control law, excepting in some singular cases, the dimension of the null space of the interaction matrix $\mathbf{L}_{sm}$ is always 2. The reason is that the 3D relationship between both image points has been lost. As a consequence, in the equilibrium state (when the $3D$ point is centered in regard with the stereovision sensor apparatus), only 2 d.o.f. are available to perform motions around the object.

We show in the experiments that an undesired rotation around the z axis is present as a side-effect, due to the wrong dimension of the null space of the interaction matrix.

3 Control

The control law used in this study is based on the *Task function* formalism [11], firstly applied to visual servoing by Espiau *et al.* [4]. In this approach, the control is directly specified in terms of regulation in the image. It may be noted that this approach has the advantage of avoiding the intermediate step of the 3D estimation of the target with regard to the end effector [9, 12]. For a given robotics task, a *target image* is built, corresponding to the desired position of the end effector with regard to the environment. If the image jacobian is not full rank (number of d.o.f > number of independent visual features), it is possible to use an hybrid task. In an hybrid task, the primary task $\mathbf{e}_1$ maintains a visual constraint during the trajectory, while the secondary task $\mathbf{e}_2$ can be seen as representing a minimization of a secondary cost h_s.

A global *task function* $\mathbf{e}$ is then defined as:

$$\mathbf{e} = \mathbf{W}^+ \mathbf{e}_1 + \gamma (\mathbb{I}_n - \mathbf{W}^+ \mathbf{W}) \frac{\partial h_s}{\partial r}^T \tag{10}$$

where $\mathbf{W}^+$ and $(\mathbb{I}_n - \mathbf{W}^+\mathbf{W})$ are two projection operators which guarantee that the camera motions due to the secondary task are compatible with the regulation of $\mathbf{s}$ to $\mathbf{s}^*$. $\mathbf{W}$ is a full rank matrix *with the same null space* as that of the interaction matrix. The parameter γ is used to tune the preponderance between the primary and the secondary task.

Considering a motionless environment, the control law has the following expression:

$${}^{\mathbf{v}} = -\lambda \mathbf{e} - \gamma (\mathbb{I}_n - \mathbf{W}^+ \mathbf{W}) \frac{\partial}{\partial t} \left(\frac{\partial h_s}{\partial r} \right)^T \tag{11}$$

This control law is applied to both presented modelings, where matrix $\mathbf{W}$ is defined as follows for each control law:

	Real stereo	Stacked-mono
$\mathbf{W}$	$\mathbf{L}^*_{3D}$	$\mathbf{L}^*_{sm}$

The symbol * is used to precise that the corresponding expression is evaluated at the equilibrium situation.

4 Experimental results

The stereo system consists of two NTSC color cameras mounted on the end-effector of a Mitsubishi PA-10 manipulator. Cameras are coarsely positioned, being approximately mounted with the same orientation, and at equal distances from the end-effector's origin. No calibration procedure has been used.

A video-rate color segmentation system is used which extracts colored regions from an image and delivers the coordinates of its centroid, its aspect ratio, and the orientation of its major axis of inertia.

Each camera is connected to one of such image processing systems. Though the system is capable of sustaining a 60 Hz frame rate, only even or odd frames are used, due to alignment problems with interlaced frames, thus reducing the frame rate to 30 Hz.

An overview of the stereovision system is depicted in Fig. 2.

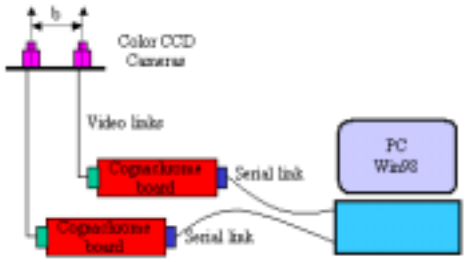

Figure 2: Overview of the stereovision system.

Though our setup is equivalent to the presented by Maru *et al.* [10], it must be noted that our task is under-constrained, and a secondary task has been defined for the motion around the object. Their object is modeled as a square and the feature vector is composed of four points (the corners).

4.1 Estimation of depth

In Fig. 1, f represents the focal length. We can write:

$$\begin{cases} sin(\alpha) = \frac{u_l}{F_u} = \frac{x_l}{z_l} \\ sin(\beta) = \frac{u_r}{F_u} = \frac{x_r}{z_r} \end{cases} \quad (12)$$

and finally:

$$x_l = x_r + b \quad (13)$$

where $F_u = \frac{f}{du}$ is the focal length along of the u axis.

With the relations 12 and 13, the depth of the observed point can be estimated as:

$$z = z_r = z_l = b.\frac{F_u}{u_l - u_r} \quad (14)$$

As a result, it is very simple to show one of the main advantages of a stereovision system in regard with a monocular vision system: the estimation of the depth. This estimation can be provided by:

$$\boxed{\hat{z} = b.\frac{Fu}{u_l - u_r}}$$

The Mitsubishi PA-10 manipulator has 7 d.o.f and is mounted on a mobile platform (XR4000 from Nomadic Inc.). In this implementation, the arm is only used and controlled as a Cartesian frame with 6 d.o.f.

The experimentation has been split in two steps. In the first step, a positioning task is executed during 300 iterations (one iteration corresponds to 33 ms). Then, the second step consists in a secondary task using a sinusoidal wave translation signal in x and y direction ($T_x = A_x.\omega_x.cos(\omega_x.t)$, $T_y = -A_y.\omega_y.sin(\omega_y.t)$ with $A_x = A_y = 0.6\ m$ and $\omega_x = \omega_y = 0.2\pi\ rd/s$). The aim of the secondary task is to describe a circle trajectory on a sphere while fixing the object centered in the image plane at a given distance.

The following table shows the different parameters (intrinsic and extrinsic) of both cameras, which have been roughly estimated:

F_u	F_v	b
300	450	118mm

The gains in the control laws are fixed to 1 for λ and 1/5 for γ.

4.2 Positioning task

In this paragraph, some results obtained in the real context when using both laws (*real stereo* and *stacked-mono*) are compared. The curves or graphs on the left side correspond to the *stacked-mono* control law, while those on the right side correspond to the *real stereo*. The reference feature to reach at the equilibrium is arbitrary fixed to $\mathbf{s}^* = (40, 0, -40, 0)^T$.

Figure 3 presents the servoing task during all the experimentation (both positioning and moving around the object).

Figure 4 gives the trajectories of the $2D$ points (left and right) in the same image plane.

Figures 5, 6, and 7, present the servoing task **only** during the positioning task. The sensor signals and the control vector have an exponential decay, but there is a persistent offset at the equilibrium. In fact, the sensor apparatus is not well calibrated, and the equilibrium sensor vector has been defined without taking into account this fact. To solve this problem, one way is to learn the desired sensor vector at the equilibrium with the uncalibrated

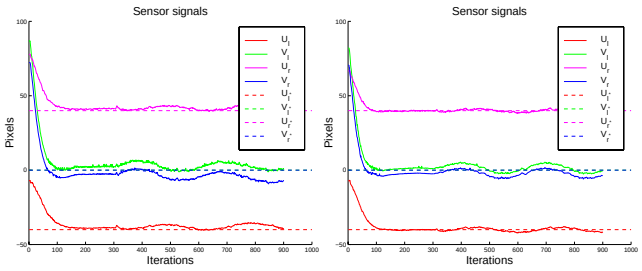

Figure 3: Sensor signals during the whole task

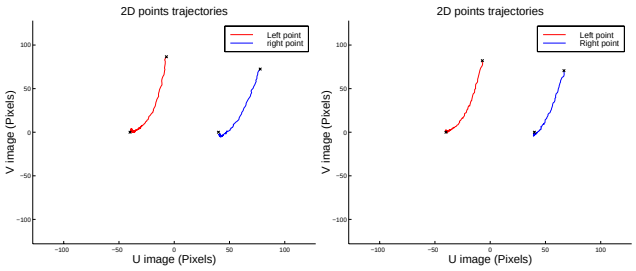

Figure 4: Image point trajectories

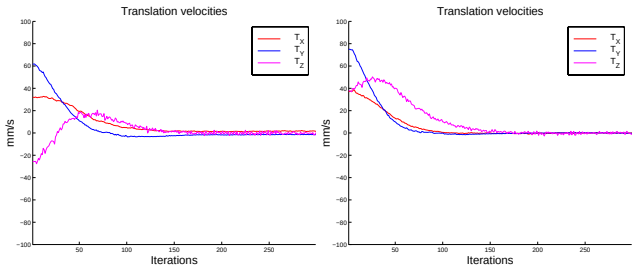

Figure 6: Translation velocities

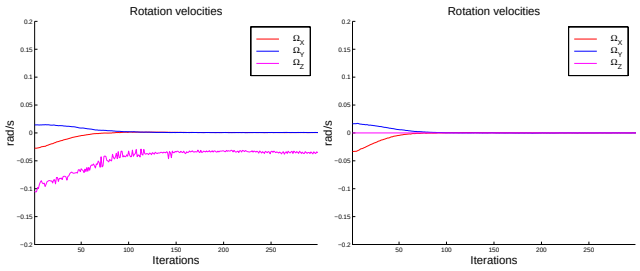

Figure 7: Rotation velocities

sensor. This is the reason why, in the *stacked-mono* approach, there exists a persistent rotation velocity in z direction (Ω_z). However, in the *real stereo*, this motion is cancelled due to the proper choice of matrix $\mathbf{W}$.

4.3 Secondary task

Figures 8, 9, and 10, present the servoing task during the secondary task.

The sensor signals do not remain in their equilibrium values: an offset due to the tracking error is present. In addition, in figures 9 and 10 the effect of the secondary task can be verified: the translation velocities T_x and T_y produce rotation velocities on Ω_y, Ω_x and a translation velocity on T_z (this corresponds to the vectors $\begin{cases} v_1 = (\ 0,\ z,\ -y,\ 1,\ 0,\ 0\ )^T \\ v_2 = (\ -z,\ 0,\ x,\ 0,\ 1,\ 0\ )^T \end{cases}$ of the kernel of the image jacobian).

Finally, in figure 10 (right side) the effect of the choice of the matrix $\mathbf{W}$ which allows to suppress the rotation velocity Ω_z, can be verified. This fact demonstrates the main advantage of the *real stereo* control.

5 Summary and Conclusions

This paper has presented for the first time the application of stereo vision in an under-constrained visual servoing task. It has been shown that problems can appear if the relationships between frames are not properly taken into account. Particularly, it concerns some uncontrolled motions which can bring the robot in its joints limits.

On the contrary, when the modeling is correctly done, the use of a $3D$ point observed by a stereo vision system is sufficient to perform motions around an unknown and complex object.

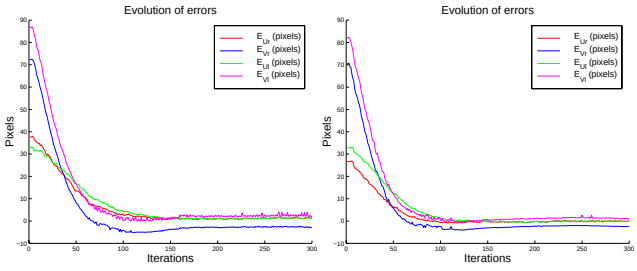

Figure 5: Evolution of the errors

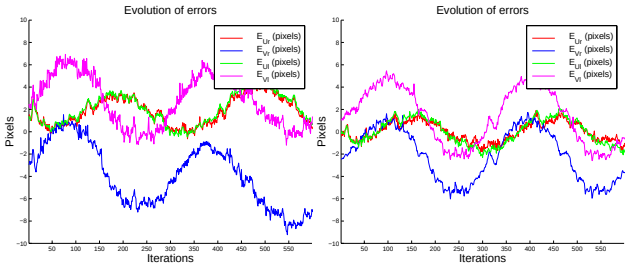

Figure 8: Evolution of the errors

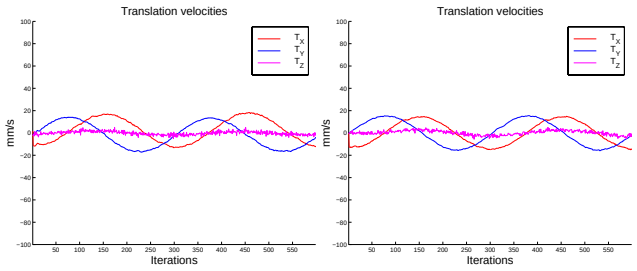

Figure 9: Translation velocities

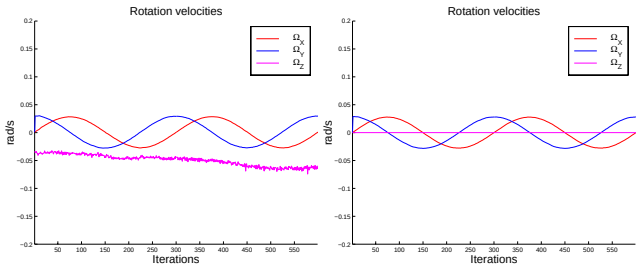

Figure 10: Rotation velocities

The choice of the centroid of the blob in both image planes is not the ideal invariant feature to perform this kind of task. As explained in [1], the center of a global bounding box is more relevant when using complex object. Future developments will concern the extension of the approach to a set of points, and other visual features (orientation and size of the blob).

First theoretical studies indicate that, when using two points in stereo, the null space of matrix $\mathbf{L}_{st}$ remains the same as that of $\mathbf{L}_{3D}$, which corresponds to a rotation around the line joining both points. With additional points, the null space is empty, thus we are interested in finding out which other properties are shared by both interaction matrices.

Acknowledgement

This work is partially funded by the Valencian Government under grant GV99-67-1-14, and a grant for a temporal stay of P. Martinet at Jaume-I University.

References

[1] F. Berry, P. Martinet, and J. Gallice. Real time visual servoing around a complex object. *IEICE Transactions on Information and Systems, Special Issue on Machine Vision Applications*, E83-D(7):1358–1368, July 2000.

[2] J. Crowley, M. Mesrabi, and F. Chaumette. Comparison of kinematic and visual servoing for fixation. In *Proc. IROS'95*, volume 1, pages 335–341, Pittsburgh, USA, 1995.

[3] K. Deguchi and Takhashi I. Image based simultaneous control of robot and target object motions by direct image interpretation method. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, volume 1, pages 375–380, Kyongju, Korea, 17-21 October 1999. IROS'99.

[4] B. Espiau, F. Chaumette, and P. Rives. A new approach to visual servoing in robotics. *IEEE Transactions on Robotics and Automation*, 8(3):313–326, 1992.

[5] G. Hager, W. C. Chang, and A. S. Morse. Robot hand-eye coordination based on a stereo vision. *IEEE Control Systems Magazine*, 15(1):30–39, 1995.

[6] M. Jägersand, O. Fuentes, and R. Nelson. Experimental evaluation od uncalibrated visual servoing for precision manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 3, pages 2874–2880, Albuquerque, USA, 1997. ICRA'97.

[7] B. Lamiroy, B. Espiau, N. Andreff, and R. Horaud. Controlling robots with two cameras: How to do it properly. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2100–2105, San Francisco, California, USA, 24-28 April 2000. ICRA'2000.

[8] E. Malis, F. Chaumette, and S. Boudet. Multi-cameras visual servoing. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 4, pages 2759–2764, San Francisco, California, USA, 24-28 April 2000. ICRA'2000.

[9] P. Martinet, J. Gallice, and D. Khadraoui. Vision based control law using 3d visual features. In *Proceedings of the Second World Automation Congress*, volume 3, pages 497–502, Montpellier, France, May 1996. ISRAM'96.

[10] N. Maru, H. Kase, S. Yamada, A. Nishikawa, and F. Miyazaki. Manipulator control by visual servoing with stereo vision. In *Proc. IROS'93*, pages 1866–1870, Yokohama, Japan, 1993.

[11] C. Samson, M. Le Borgne, and B. Espiau. *Robot Control. The task function approach.* ISBN 0-19-8538057. Clarendon Press, Oxford, 1991.

[12] W. J. Wilson, C. C. Williams Hulls, and G. S. Bell. Relative end-effector control using cartesian position based visual servoing. *IEEE Transactions on Robotics and Automation*, 12(5):684–696, October 1996.