Stereo Visual Servoing with a Single Point: a Comparative Study

E. Cervera

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

F. Berry, P. Martinet

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

Abstract

A comparative study of a stereo visual servoing system is presented. The visual feature is the pair of images of an object, for which there not exists a geometrical model. Instead, the image of the object is segmented, and its center of gravity is computed. The developed control laws use either the raw image points, or the estimated 3D coordinates. The stereo rig is roughly calibrated with the nominal values of the camera parameters. Experimental results on a real arm with a stereo pair mounted on the end-effector are shown. These results confirm the equivalence between the image pairs and the 3D estimation, and highlight the importance of the choice of coordinate frames in the servoing task.

1 Introduction

Stereovision information has been commonly considered as an alternative way to recover the depth, in the modeling phase of a vision system. Recent works [3] [4] have awaken new interest in stereo visual servoing application, considering mainly the robustness and precision aspects. Our goal is to analyse the case of a 3D point, to develop different model and control laws, and of course to validate experimentally those approaches.

In this work we present some theoretical results concerning the modeling of a positioning task using only one 3D point. We have tested many approaches to define this kind of task when using a stereovision system. However, the observed object is not an ideal point: its shape might be irregular and its geometrical model is not known. The object is segmented from the scene by the vision system, and its center of gravity is computed. The image coordinates of this point in both cameras are the output of the vision system to the robot controller.

First, we consider the modeling of the two stereo images of the point. In this case, care must be taken with the definition of the coordinates frame of the cameras and the end-effector. This is the so-called *image based approach* [2]. Second, we have developed and tested different models of a 3D point feature, either using the estimated Cartesian coordinates, or combining the image coordinates with the estimated depth of the 3D point.

It should be noted that the only source of information is the stereo rig. Thus, all the 3D information is estimated from these measurements, as well as from the intrinsic and extrinsic camera parameters (which are roughly known). Our interest is to compare the approaches to test whether there exists an advantage in using either the raw signals or the computed 3D features.

2 Task definition

Our setup consists of a stereo rig mounted on the end-effector of the manipulator.

Let us define $\mathcal{F}_e$ as the control 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.

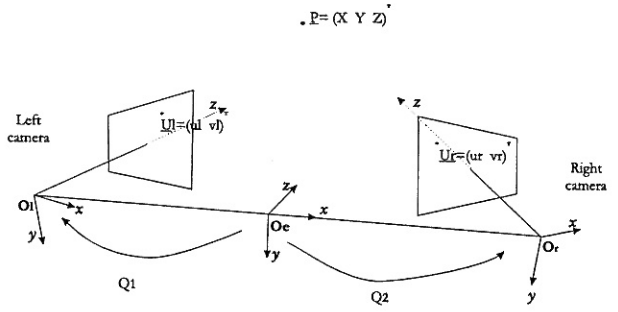

Figure 1: Configuration of a general stereo vision system.

The raw feature vector is defined as

$$\underline{s} = (u_l, v_l, u_r, v_r)^T \quad (1)$$

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.

In a general case, the cameras are not aligned to the control frame $\mathcal{F}_e$. In our case, the features computed from the visual features $\underline{s}$ is the position of the spatial point $\underline{P}$ in the frame $\mathcal{F}_e$. From this visual feature, we propose several control laws with a comparative study. At first, let us express the coordinate of $\underline{P}$ in function of $\underline{s}$. To compute the location of $\underline{P}$, we define two homogenous transformation matrices Q_1 and Q_2 such as

$$Q_1 \colon \mathcal{F}_e \rightarrow \mathcal{F}_l$$

$$Q_2 \colon \mathcal{F}_e \rightarrow \mathcal{F}_r$$

These homogenous transformations are supposed to be known (or evaluated) and can be written as follow

$$Q_i = \begin{pmatrix} & R_i & & T_i \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where R and T are the rotation and the translation of the transformation respectively.

Two others transformations are necessary to project the point from the camera frame $\mathcal{F}_l$ and $\mathcal{F}_r$ to the image space $\mathcal{I}_l$ and $\mathcal{I}_r$. These transformations denoted C_1 and C_2 are defined as

$$C_1 \colon \mathcal{F}_l \rightarrow \mathcal{I}_l$$

$$C_2 \colon \mathcal{F}_r \rightarrow \mathcal{I}_r$$

These transformations are composed by the intrinsec parameters of the cameras and can be written as follows

$$C_i = \begin{pmatrix} F_u & \theta_{uv} & u_0 \\ 0 & F_v & u_0 \\ 0 & 0 & 1 \end{pmatrix}$$

where i=1 for the left camera and i=2 for the right. F_u, F_v are the focal length along x and y, θ_{uv} takes into account the angle between the axis x and y, and $(u_0\ v_0)^T$ are the coordinates of the optical center. So, the image point $\underline{U_l}$ can be easily computed from $\underline{P}$ expressed in $\mathcal{F}_e$

$$\underline{U_l} = C_1 . \begin{pmatrix} R_1 & T_1 \end{pmatrix} . \underline{P}$$

This relationship can be rewritten under a global matrix such as

$$\begin{pmatrix} s.u_l \\ s.v_l \\ s \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{pmatrix} . \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \quad (2)$$

where s is the scale factor, and m_{ij} are the elements of the transformation $C_1.Q_1$. For the right camera, the same approach gives

$$\begin{pmatrix} s.u_r \\ s.v_r \\ s \end{pmatrix} = \begin{pmatrix} m'_{11} & m'_{12} & m'_{13} & m'_{14} \\ m'_{21} & m'_{22} & m'_{23} & m'_{24} \\ m'_{31} & m'_{32} & m'_{33} & m'_{34} \end{pmatrix} . \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \quad (3)$$

where m'_{ij} are the elements of the transformation $C_2.Q_2$.

A development of the relations (2) and (3) gives:

$$\begin{cases} u_l = \frac{m_{11}.X+m_{12}.Y+m_{13}.Z+m_{14}}{m_{31}.X+m_{32}.Y+m_{33}.Z+m_{34}} \\ v_l = \frac{m_{21}.X+m_{22}.Y+m_{23}.Z+m_{24}}{m_{31}.X+m_{32}.Y+m_{33}.Z+m_{34}} \\ u_r = \frac{m'_{11}.X+m'_{12}.Y+m'_{13}.Z+m'_{14}}{m'_{31}.X+m'_{32}.Y+m'_{33}.Z+m'_{34}} \\ v_r = \frac{m'_{21}.X+m'_{22}.Y+m'_{23}.Z+m'_{24}}{m'_{31}.X+m'_{32}.Y+m'_{33}.Z+m'_{34}} \end{cases} \quad (4)$$

and the resolution of this system of 4 equations allows to solve the position of the point $\underline{P} = \begin{pmatrix} X & Y & Z \end{pmatrix}^T$.

In our case, we consider a simplified configuration where both cameras are parallel with identical focal lengths (F_u, F_v) and the control frame $\mathcal{F}_e$ is located at the center of both frames (Fig 2). Both cameras are aligned along the x-axis and the distance between them is b.

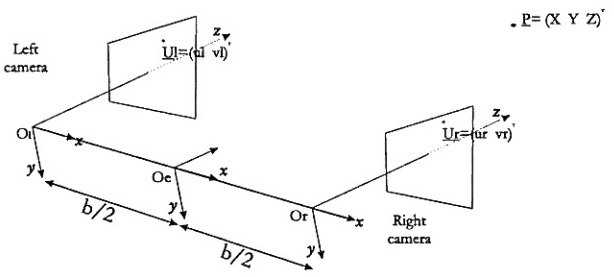

Figure 2: The simplified configuration of our system.

Thus, the system (4) becomes

$$\begin{cases} u_l = \frac{F_u.X + F_u.b/2}{Z} \\ v_l = \frac{F_v.Y}{Z} \\ u_r = \frac{F_u.X - F_u.b/2}{Z} \\ v_r = \frac{F_v.Y}{Z} \end{cases} \quad (5)$$

and the coordinates of the observed point in $\mathcal{F}_e$ can be easily deduced as

$$\underline{\hat{P}} = \begin{pmatrix} \hat{X} \\ \hat{Y} \\ \hat{Z} \end{pmatrix} = \begin{pmatrix} \frac{b.(u_l+u_r)}{2.(u_l-u_r)} \\ \frac{b.v_r.F_u}{(u_l-u_r).F_v} \\ \frac{b.F_u}{u_l-u_r} \end{pmatrix} \quad (6)$$

These values are roughly estimated or are taken directly from their nominal values. No explicit calibration procedure has been undertaken.

Again, it should be noted that this is not a physical point. The image coordinates correspond to the center of gravity of each segmented object. Since the shape observed by each camera is slightly different, the 3D point is only an abstract point which only in ideal cases would correspond to the real center of gravity of the object. However, the robustness of the visual servoing approach makes it possible to use these features in the control loop.

3 Control laws

Depending on the chosen feature vector, different control laws can be used. If the raw image measurements are used, we obtain two *stereo* control laws. The difference between them relies on whether the transformation between the end-effector frame and the camera frames is taken into account or not.

3.1 Stereo 2D point

In this case, this transformation is taken into account explicitly in the control law.

The feature vector is the raw image information (Eq. 1) and the jacobian matrix is

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_l {}^l\mathbf{M}_e \\ \mathbf{L}_r {}^r\mathbf{M}_e \end{pmatrix} \quad (7)$$

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

$$\mathbf{L}_i = \begin{pmatrix} -\frac{F_u}{z} & 0 & \frac{u_i}{z} & \frac{u_i v_i}{F_v} & -F_u - \frac{u_i^2}{F_u} & \frac{v_i F_u}{F_v} \\ 0 & -\frac{F_u}{z} & \frac{v_i}{z} & F_v + \frac{v_i^2}{F_v} & -\frac{u_i v_i}{F_u} & -\frac{u_i F_v}{F_u} \end{pmatrix} \quad (8)$$

and ${}^l\mathbf{M}_e$ 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 kinematic screws $\mathbf{v}$ is

$${}^i\mathbf{v} = {}^i\mathbf{M}_e {}^e\mathbf{v} \quad (9)$$

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

$${}^i\mathbf{M}_e = \begin{pmatrix} {}^i\mathbf{R}_e & [{}^i\mathbf{t}_e]_\times {}^i\mathbf{R}_e \\ \mathbf{O}_3 & {}^i\mathbf{R}_e \end{pmatrix} \quad (10)$$

It can be shown that the resulting interaction matrix (Eq. 7) is the same as that obtained by Maru *et al.* [6].

3.2 Simplified stereo 2D point

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 stereovision, 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.

The raw feature vector is used as before but the jacobian matrix is now

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_l \\ \mathbf{L}_r \end{pmatrix} \quad (11)$$

3.3 3D point

Since the 3D coordinates of the observed point can be computed from the image data (and an estimation of the extrinsic and intrinsic parameters of the cameras), they can also be used in the control law.

Thus, the feature vector is composed by the estimated coordinates (Eq. 6) and the jacobian matrix is

$$\mathbf{L} = \begin{pmatrix} -\mathbf{I}_3 & \left[\underline{\hat{P}}\right]_\times \end{pmatrix} \quad (12)$$

3.4 Linearized 2D point and Z

Cervera and Martinet [1] have shown that a feature vector where the image coordinates are multiplied by z is equivalent to the estimation of the real 3D coordinates. With this approach, the feature vector is

$$\underline{\mathbf{s}} = (z.u_r, z.v_r, z)^T \quad (13)$$

and the jacobian matrix is

$$\mathbf{L} = \begin{pmatrix} F_u & 0 & 0 \\ 0 & F_v & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -\mathbf{I}_3 & \left[\underline{\hat{P}}\right]_\times \end{pmatrix} \quad (14)$$

3.5 Raw 2D point and Z

Finally, the combination of an image point and the estimation of z can be used to obtain the feature vector

$$\underline{s} = (u_r, v_r, z)^T \tag{15}$$

Depending on how the frames are managed, and how the jacobian of z is computed, there are three different alternatives.

3.5.1 End-effector frame

If the transformation between the end-effector and the camera frame is taken into account, the jacobian matrix is

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_r {}^r\mathbf{M}_e \\ -b\frac{F_u}{(u_l-u_r)^2}(\mathbf{L}_{u_l} {}^l\mathbf{M}_e - \mathbf{L}_{u_r} {}^r\mathbf{M}_e) \end{pmatrix} \tag{16}$$

where $\mathbf{L}_r$ is computed as in 8 and $\mathbf{L}_{u_r}$ and $\mathbf{L}_{u_l}$ are the first row of matrix 8 for the right and left image points respectively.

3.5.2 Camera frame

If the above-mentioned transformation is neglected, the jacobian matrix is

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_r \\ -b\frac{F_u}{(u_l-u_r)^2}(\mathbf{L}_{u_l} - \mathbf{L}_{u_r}) \end{pmatrix} \tag{17}$$

where the components of the matrix are defined as in the previous section.

3.5.3 Direct combination

Finally, the jacobian of the image point is directly combined with the third row (z component) of the jacobian of the 3D point (Eq. 12), thus obtaining

$$\mathbf{L} = \begin{pmatrix} -\frac{F_u}{z} & 0 & \frac{u_i}{z} & \frac{u_i v_i}{F_v} & -F_u - \frac{u_i^2}{F_u} & \frac{v_i F_u}{F_v} \\ 0 & -\frac{F_v}{z} & \frac{v_i}{z} & F_v + \frac{v_i^2}{F_v} & -\frac{u_i v_i}{F_u} & -\frac{u_i F_v}{F_u} \\ 0 & 0 & -1 & -z\frac{v_r}{F_v} & z\frac{u_r}{f_u} & 0 \end{pmatrix} \tag{18}$$

4 Experimental results

The mobile manipulator of the Robotic Intelligence Lab is composed by a Nomad XR4000 platform and a Mitsubishi PA-10 arm (Fig. 3). Attached to the end-effector of the arm is a stereo rig with two color cameras, linked to two video boards which deliver the visual features (center of gravity of a color-segmented object) at video rate.

A manipulation task is split in two steps. In the first one, 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 $A_x = A_y = 0.6\ m$ et $\omega_x = \omega_y = \pi/5\ rad.s^{-1}$). The aim of the secondary task is to fix at a given distance the object centered in the image plane when describing a circle trajectory on a sphere.

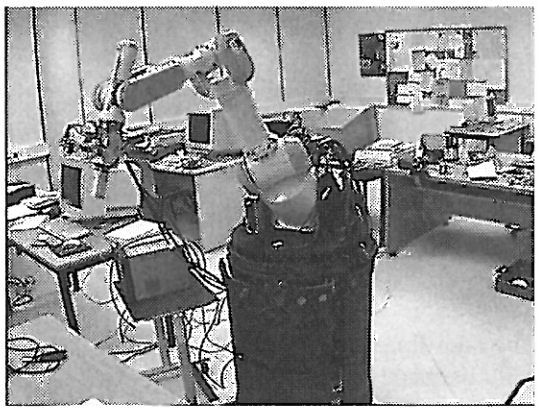

Figure 3: The stereo visual servoing manipulator setup.

The following table gives the different parameters (intrinsic and extrinsic) of both cameras.

F_u	F_v	b
300	450	118mm

The gain in the control laws is fixed to 1 for the primary task (positioning) and 0.2 for the secondary task.

4.1 Comparison

Experimental results are depicted in Figures 4 to 10. Each figure is divided into six plots which show the image errors, the estimation of Z, the translational and rotational velocities, the trajectory of the end-effector, and the image points.

The main conclusion from these experiments is the strong similarity between all the approaches, except one of them: the *simplified stereo 2D point*. In this case, the system is unstable, due to the calibration errors, which produce a rotation around the Z-axis. As shown in Fig. 5, though the image points converge to the desired values, the trajectory of the end-effector is not stable, specially during the secondary task.

This rotation is properly eliminated in all the approaches which use an estimation of Z. But, more interestingly, it is also eliminated if the transformation

between the camera and control frames is correctly taken into account, as in the *stereo 2D point.*

4.1.1 Precision

For comparison purposes, the error is directly measured on the image features. The desired position is fixed to the image point coordinates $(-40, 0)$ and $(40, 0)$ for the left and right camera respectively. This would be the final position in a precisely calibrated system. Since the real system is only roughly calibrated (i.e. the cameras are not prefectly aligned) a small error persists in components v_l and v_r.

During the secondary task, a tracking error appears, but it should be noted that the control loop is purely proportional. Consequently the error could be reduced if a higher gain were used, but stability problems could arise.

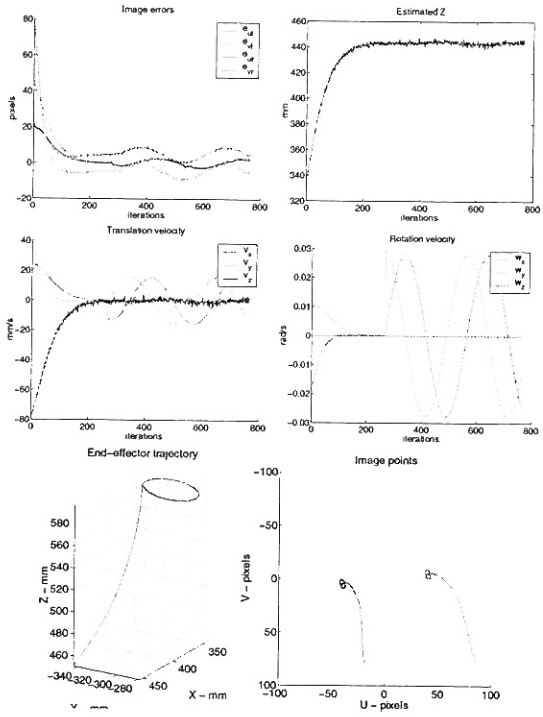

Figure 4: Stereo 2D point

4.1.2 Robustness

The dynamic response of the servoing system is depicted using the translation and rotation velocities. During the convergence phase, all the velocities decrease to zero, except for the *simplified stereo* case,

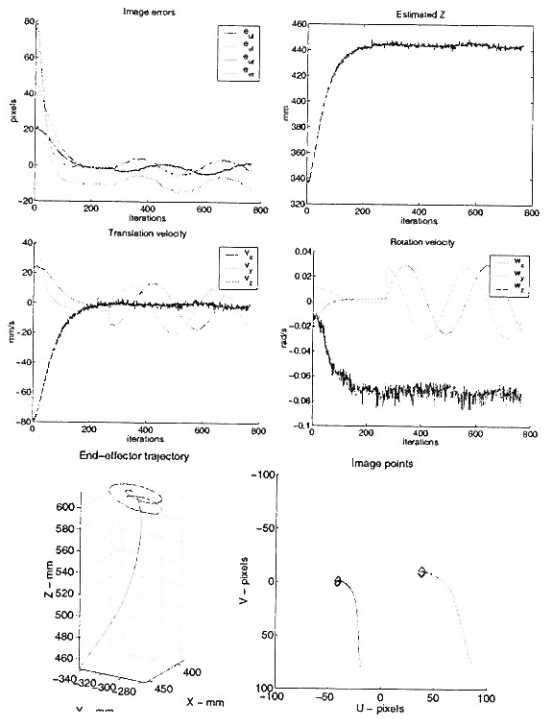

Figure 5: Simplified stereo 2D point

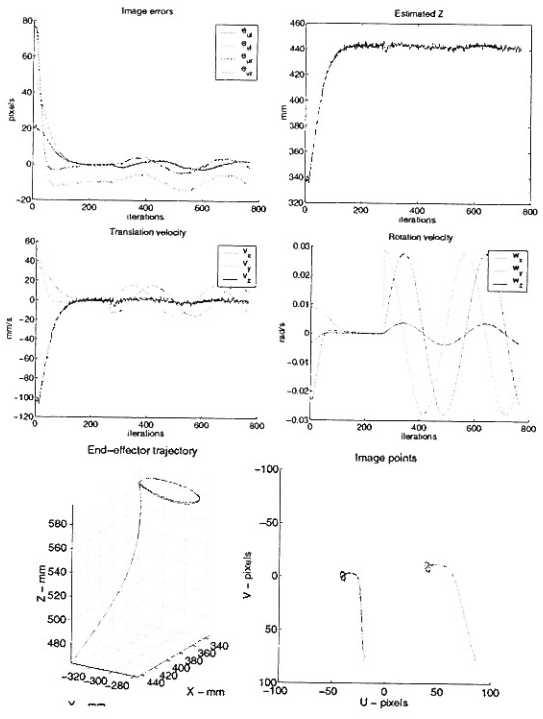

Figure 6: Estimated 3D point

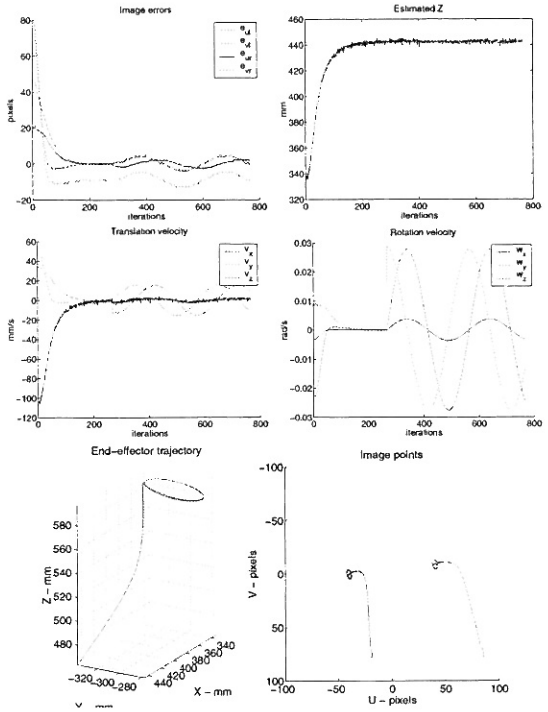

Figure 7: Linearized 2D point and Z

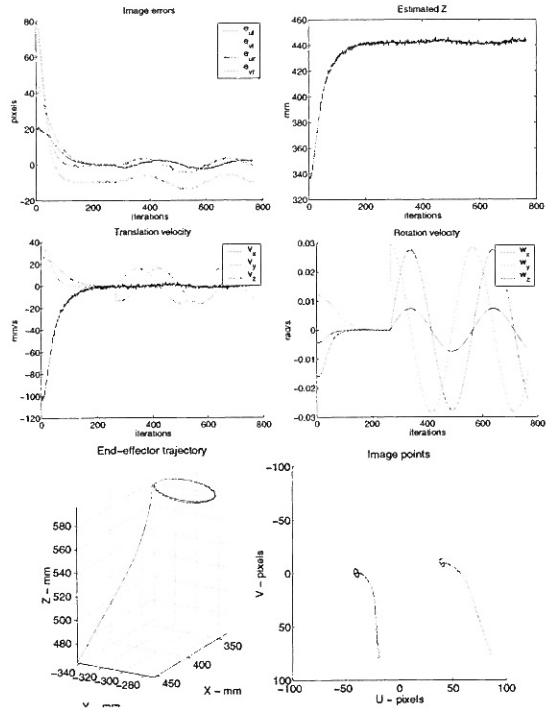

Figure 8: Raw 2D point and Z (end-effector)

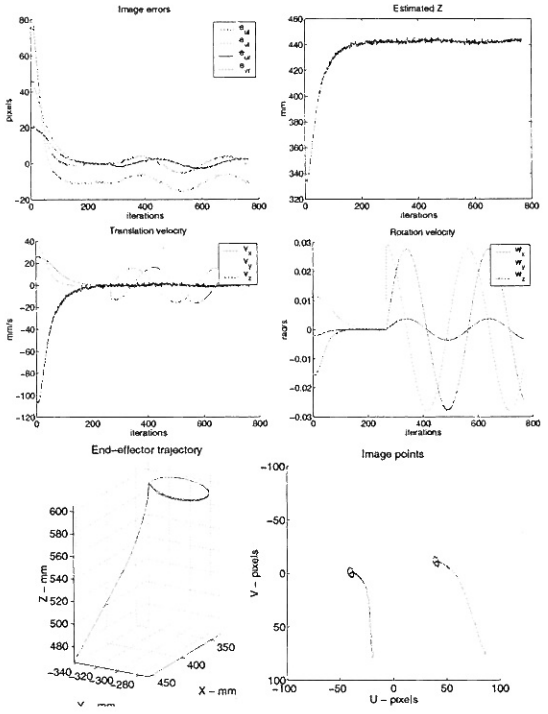

Figure 9: Raw 2D point and Z (camera)

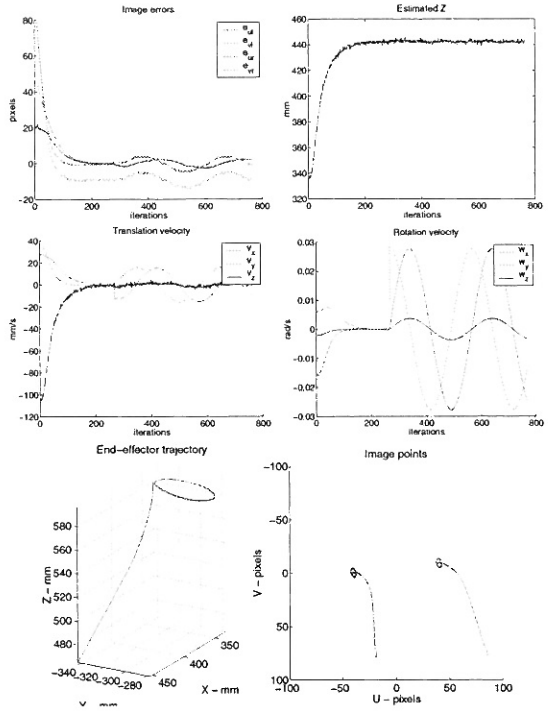

Figure 10: Raw 2D point and Z (direct)

where ω_z is not null, due to the wrong null space of the jacobian (see [5] for a more detailed study).

During the secondary task, the coupling between the translational and rotational velocities, corresponding to axes X and Y, is perfectly shown. This coupling is theoretically demonstrated by the computation of the null space of the jacobian.

However, a small oscillatory motion is present also in axis Z. Though it does not compromise the stability of the system, additional studies are required in order to identify the source of this perturbation.

5 Summary and Conclusions

In this paper we have developed different models and control laws for a stereo visual servoing system with a single point feature. The observed object is not an ideal point but an irregular-shaped object with unknown geometrical model. The object is segmented from the scene by the vision system, and its center of gravity is computed. The image coordinates of this point in both cameras is the output of the vision system to the robot controller.

The resulting system is shown to be robust against calibration errors, either when the raw image features are used or when the 3D point coordinates is estimated.

However, care must be taken in the choice of the frames. In the stereo image-based approach, if the transformation matrices are not properly taken into account, there exists a persistent rotation velocity around z direction, due to calibration error. However, if those matrices are introduced in the model, the system is robust against this error.

In the future, additional image points as well as new features like orientation and size of the segmented object are planned. These extensions should allow to perform more advanced tasks like stereo visual grasping and object recognition.

Acknowledgement

This work is partially funded by the Valencian Government (Conselleria de Cultura i Educació) under grants GV99-67-1-14 and INV00-14-61.

References

[1] E. Cervera and P. Martinet. Combining pixel and depth information in image-based visual servoing. In *Proceedings of the International Conference on Advanced Robotics*, volume 1, Tokyo, Japan, 25-27 October 1999. ICAR'99.

[2] 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.

[3] 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.

[4] 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.

[5] P. Martinet and E. Cervera. Stacking jacobians properly in stereo visual servoing. In *Proceedings of the IEEE International Conference on Robotics and Automation*, Seoul, Korea, May 2001. ICRA'01.

[6] 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.