Proceedings of the 2002 IEEE
International Conference on Robotics & Automation
Washington, DC • May 2002

Is 3D useful in stereo visual control?

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

The main goal of this paper is the study of image-based stereo visual servoing. A pair of cameras is mounted on the end-effector of the manipulator arm. The visual features are the pair of images of an unknown object. The developed control laws use either the raw image points, or the estimated 3D coordinates. The experimental setup is challenging: large rotations are involved, images are noisy, and cameras are coarsely calibrated. In this setup, the trajectory of the end-effector differs notably, sometimes leading the arm near its joint range limits. Experimental results demonstrate that using pixel coordinates is disadvantageous, compared with 3D coordinates estimated from the same pixel data.

1 Introduction

Usually, stereo visual features have been considered as an alternative way to recover the depth, in the modeling phase of a vision system. The application of stereo vision in visual servoing was pioneered by Maru *et al.* [9], and recently adressed in [1] [5] [6] [7]. Works proposed in [8] have awaken new interests, considering mainly the robustness and precision aspects. In [2], a comparative study of a stereo visual servoing system was initialized. Stereo visual servoing offers some advantages over the classical monocular 2D and 3D visual servoing approaches.

Depth information can be recovered without need of any geometrical model of the observed object. It should be noted that even in 2D visual servoing, this information is needed for the computation of the image jacobian.

As pointed out in [8], a number of singularities exists in monocular visual servoing, making visual control impossible near those configurations. These singularities can be avoided by using a stereo rig, thus requiring less strict camera calibration.

This paper presents a visual servoing approach based on stereo vision. We experimentally show that using 3D coordinates (estimated from the stereo images) in the feature vector performs better than using raw 2D image coordinates. In our experimental setup, the stereo rig is mounted on the end-effector of the arm. The programmed manipulation task is quite challenging: large rotations are involved, pixel noise is high, and camera calibration is coarse.

The rest of this paper is organized as follows: first, we consider the modeling of two stereo images of a set of points, both in the general case, and in the simplified case where cameras are aligned.

Next, we develop visual control with three different features: in the first one, raw pixel coordinates are used. This is the so-called *image based approach* [4]. Care must be taken with the definition of the coordinates frame of the cameras and the end-effector.

Image-based 3D features are then introduced: estimated coordinates, and a combination of pixel data and stereo disparity. We show that this third approach exhibits the same nice properties as using coordinates, with regard to the end-effector trajectory.

Finally, we present experimental results of the presented approaches, with a comparison of image feature errors, the velocity screw, and the trajectory of the end-effector.

It should be noted that, in all of the approaches, 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 Stereo Observation of a Set of Points

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

0-7803-7272-7/02/$17.00 © 2002 IEEE

to the right camera.

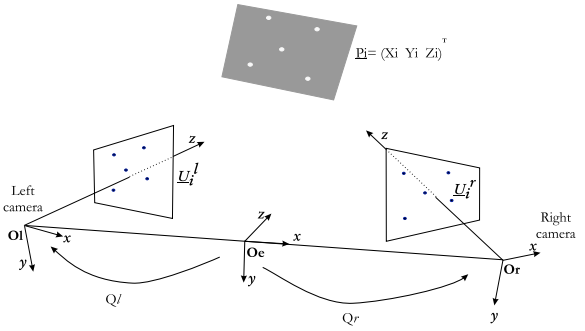


Figure 1: Configuration of a general stereo vision system.

In this work, a segmented target defined by 5 points is considered. The corresponding raw feature vector is defined by

$$\underline{\mathbf{s}} = \left[u_1^l, v_1^l, u_1^r, v_1^r, \cdots, u_5^l, v_5^l, u_5^r, v_5^r\right]^T \tag{1}$$

where $\underline{\mathbf{U}}_i^l = (u_i^l, v_i^l)^T$ and $\underline{\mathbf{U}}_i^r = (u_i^r, v_i^r)^T$ are the image coordinates of the i^{th} point, observed by the left and right cameras respectively and $\underline{\mathbf{s}}_i$ is the ith subvector of $\underline{\mathbf{s}}$ such $\underline{\mathbf{s}}_i = (\underline{\mathbf{U}}_i^l \;\; \underline{\mathbf{U}}_i^r)$.

In a general case, the cameras are not aligned with the control frame $\mathcal{F}_e$ (Fig. 1). Coordinates of spatial points $\underline{\mathbf{P}}_i$ (i=1$\cdots$5) in frame $\mathcal{F}_e$ can be computed from visual data $\underline{\mathbf{s}}_i$. From this visual feature, we propose several control laws with a comparative study. At first, let us express coordinates of $\underline{\mathbf{P}}_i$ in function of $\underline{\mathbf{s}}_i$. To compute the location of $\underline{\mathbf{P}}_i$, we define two homogenous transformation matrices Q_l and Q_r such as

$$Q_l\colon \mathcal{F}_e \rightarrow \mathcal{F}_l$$
$$Q_r\colon \mathcal{F}_e \rightarrow \mathcal{F}_r$$

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

$$Q_{l \text{ or } r} = \begin{pmatrix} & R & & T \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where R and T is 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_l and C_r are defined as

$$C_l\colon \mathcal{F}_l \rightarrow \mathcal{I}_l$$
$$C_r\colon \mathcal{F}_r \rightarrow \mathcal{I}_r$$

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

$$C_{l \text{ or } r} = \begin{pmatrix} F_u & \theta_{uv} & u_0 \\ 0 & F_v & u_0 \\ 0 & 0 & 1 \end{pmatrix}$$

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 $\underline{\mathbf{U}}_i^l$ can be easily computed from $\underline{\mathbf{P}}_i$ expressed in $\mathcal{F}_e$

$$\underline{\mathbf{U}}_i^l = C_l.Q_l.\underline{\mathbf{P}}_i$$

This relationship can be rewritten under a global matrix such as

$$\begin{pmatrix} u_i^l\alpha \\ v_i^l\alpha \\ \alpha \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{pmatrix} \cdot \begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix} \tag{2}$$

where α is the scale factor, and m_{ij} are the elements of the transformation C_lQ_l. For the right camera, the same approach gives

$$\begin{pmatrix} u_i^r\alpha \\ v_i^r\alpha \\ \alpha \end{pmatrix} = \begin{pmatrix} m'_{11} & m'_{12} & m'_{13} & m'_{14} \\ m'_{21} & m'_{22} & m'_{23} & m'_{24} \\ m'_{31} & m'_{32} & m'_{33} & m'_{34} \end{pmatrix} \cdot \begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix} \tag{3}$$

where m'_{ij} are the elements of the transformation C_rQ_r.

A development of relationships (2) and (3) gives:

$$\begin{cases} u_i^l = \frac{m_{11}X_i+m_{12}Y_i+m_{13}Z_i+m_{14}}{m_{31}X_i+m_{32}Y_i+m_{33}Z_i+m_{34}} \\ v_i^l = \frac{m_{21}X_i+m_{22}Y_i+m_{23}Z_i+m_{24}}{m_{31}X_i+m_{32}Y_i+m_{33}Z_i+m_{34}} \\ u_i^r = \frac{m'_{11}X_i+m'_{12}Y_i+m'_{13}Z_i+m'_{14}}{m'_{31}X_i+m'_{32}Y_i+m'_{33}Z_i+m'_{34}} \\ v_i^r = \frac{m'_{21}X_i+m'_{22}Y_i+m'_{23}Z_i+m'_{24}}{m'_{31}X_i+m'_{32}Y_i+m'_{33}Z_i+m'_{34}} \end{cases} \tag{4}$$

and the resolution of this system of four equations allows to solve the position of the i^{th} point $\underline{\mathbf{P}}_i = (X_i, Y_i, Z_i)^T$.

In our case, we consider a simplified configuration where the 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 the both frames (Fig 2). Both cameras are aligned along the x-axis and the distance between them is b.

Thus, the system (4) becomes

$$\begin{cases} u_i^l &= \frac{F_uX_i+F_ub/2}{Z_i} + u_0 \\ v_i^l &= \frac{F_vY_i}{Z_i} + v_0 \\ u_i^r &= \frac{F_uX_i-F_ub/2}{Z_i} + u_0 \\ v_i^r &= \frac{F_vY_i}{Z_i} + v_0 \end{cases} \tag{5}$$

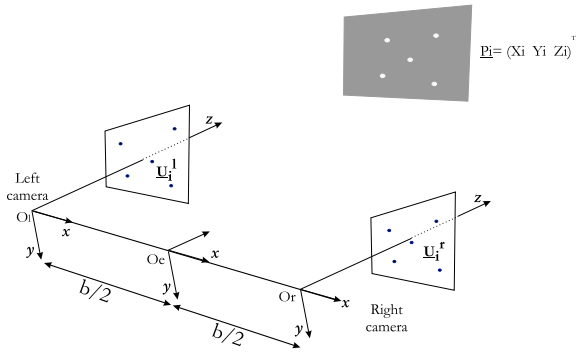


Figure 2: The simplified configuration of our system.

and the coordinates of the observed point can be easily deduced as

$$\underline{\hat{\mathbf{P}}}_i = \begin{pmatrix} \hat{X}_i \\ \hat{Y}_i \\ \hat{Z}_i \end{pmatrix} = \begin{pmatrix} \frac{b(u_i^l+u_i^r-2u_0)}{2(u_i^l-u_i^r)} \\ \frac{bF_u(v_i^r+v_i^l-2v_0)}{2(u_i^l-u_i^r)F_v} \\ \frac{bF_u}{u_i^l-u_i^r} \end{pmatrix} \tag{6}$$

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

3 Visual Features

The essence of visual servoing is the computation of the matrix of derivatives (the jacobian) of the visual feature vector with respect to the velocity screw. Using the raw pixel data or the estimated 3D point coordinates is a matter of choice. Both approaches require an estimation of camera parameters. However, the resulting dynamic properties of the task may differ. In this section, we present the theoretical bases of both approaches, and a third feature vector which uses the stereo disparity, without fully estimating the real 3D coordinates.

3.1 Stereo 2D point

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

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_1^l \mathbf{M}_e^l \\ \mathbf{L}_1^r \mathbf{M}_e^r \\ \vdots \\ \mathbf{L}_5^l \mathbf{M}_e^l \\ \mathbf{L}_5^r \mathbf{M}_e^r \end{pmatrix} \tag{7}$$

where $\mathbf{L}_i^l$ and $\mathbf{L}_i^r$ are the interaction matrices for i^{th} point, relative to the left and right cameras respectively, as defined by Espiau *et al.* [4]

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

The dimension of the final image Jacobian $\mathbf{L}$ is 20×6. $\mathbf{M}_e^l$ and $\mathbf{M}_e^r$ 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 $\mathcal{F}_j$, the relationship between the kinematic screws $\mathbf{v}$ is

$$\underline{\mathbf{v}}^j = \mathbf{M}_e^j \underline{\mathbf{v}}^e \tag{9}$$

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

$$\mathbf{M}_e^j = \begin{pmatrix} \mathbf{R}_e^j & [\underline{\mathbf{t}}_e^j]_\times \mathbf{R}_e^j \\ \mathbf{O}_3 & \mathbf{R}_e^j \end{pmatrix} \tag{10}$$

It can be shown that the resulting interaction matrix (7) is the same as that obtained by Maru *et al.* [9].

3.2 2D points and disparity

Since the 2D image features, and the stereo disparity of the i^{th}point $(u_i^l - u_i^r)$ can be computed from the image data. In the following control law, the feature vector is defined as

$$\begin{aligned} \underline{\mathbf{s}} = \Big(& \tfrac{u_1^l+u_1^r}{u_1^l-u_1^r}, \tfrac{v_1^l+v_1^r}{u_1^l-u_1^r}, \tfrac{1}{u_1^l-u_1^r} \cdots \\ & \cdots \tfrac{u_5^l+u_5^r}{u_5^l-u_5^r}, \tfrac{v_5^l+v_5^r}{u_5^l-u_5^r}, \tfrac{1}{u_5^l-u_5^r} \Big)^T \end{aligned} \tag{11}$$

It can be shown that this vector results from a linear combination of the 3D coordinates of the corresponding 3D point:

$$\underline{\mathbf{s}} = \left((A\underline{\hat{\mathbf{P}}}_1)^T \cdots (A\underline{\hat{\mathbf{P}}}_5)^T \right)^T \tag{12}$$

where

$$A = \begin{pmatrix} \frac{2}{b} & 0 & \frac{2u_0}{bF_u} \\ 0 & \frac{2F_v}{bF_u} & \frac{2v_0}{bF_u} \\ 0 & 0 & \frac{1}{bF_u} \end{pmatrix}$$

and the resulting Jacobian matrix for one point is as shown in Equation 13.

The interest in using this model is twofold: the 3D coordinates need not to be estimated, and the jacobian matrix is linear with respect to $\underline{\mathbf{s}}$. Effectively, as shown in [3], the jacobian matrix (13) can be expressed as

$$\mathbf{L} = \begin{pmatrix} -\mathbf{A} & \mathbf{A}\left[\mathbf{A}^{-1}\underline{\mathbf{s}}_1\right]_\times \\ \vdots & \\ -\mathbf{A} & \mathbf{A}\left[\mathbf{A}^{-1}\underline{\mathbf{s}}_5\right]_\times \end{pmatrix} \tag{14}$$

$$\mathbf{L}_i = \begin{pmatrix} -\frac{2}{b} & 0 & -\frac{2u_0}{bF_u} & -\frac{u_0(v_i^l+v_i^r-2v_0)}{F_v(u_i^l-u_i^r)} & \frac{u_0(u_i^l+u_i^r-2u_0)-2F_u^2}{F_u(u_i^l-u_i^r)} & \frac{F_u(v_i^l+v_i^r-2v_0)}{F_v(u_i^l-u_i^r)} \\ 0 & -\frac{2F_u}{bF_u} & -\frac{2v_0}{bF_u} & -\frac{v_0(v_i^l+v_i^r-2v_0)-2F_v^2}{F_v(u_i^l-u_i^r)} & \frac{v_0(u_i^l+u_i^r-2u_0)}{F_u(u_i^l-u_i^r)} & -\frac{F_v(u_i^l+u_i^r-2u_0)}{F_u(u_i^l-u_i^r)} \\ 0 & 0 & -\frac{1}{bF_u} & -\frac{v_i^l+v_i^r-2v_0}{2F_v(u_i^l-u_i^r)} & \frac{u_i^l+u_i^r-2u_0}{2Fu(u_i^l-u_i^r)} & 0 \end{pmatrix} \quad (13)$$

where $\underline{\mathbf{s}}_i$ is the ith element of $\underline{\mathbf{s}}$ such $\underline{\mathbf{s}}_i = (A\hat{\underline{\mathbf{P}}}_i)^T$. Additionally, some theoretical results from 3D points still hold for any linear combination: though the velocity screw (Eq. 16) is valid for small angles only, the trajectory of the center of gravity of the set of points *still translates along a straight path* during the task (see [3] for details).

3.3 Estimated 3D point

Instead of using the 2D coordinates of the observed point, we have experimented with the estimated 3D coordinates. Thus, the feature vector consists of the estimated coordinates (eq. 6) and the jacobian matrix is

$$\mathbf{L} = \begin{pmatrix} -\mathbf{I}_3 & \left[\hat{\underline{\mathbf{P}}}_1\right]_\times \\ & \vdots \\ -\mathbf{I}_3 & \left[\hat{\underline{\mathbf{P}}}_5\right]_\times \end{pmatrix} \quad (15)$$

The main advantage of using 3D features is the linearity of the jacobian matrix. As a result, some theoretical properties of the trajectory of the end-effector can be obtained. Effectively, Cervera and Martinet [3] demonstrated, for a feature vector composed of a rather general set of 3D points, that the velocity screw of the camera is

$$\underline{\mathbf{v}} = -\lambda \begin{bmatrix} (\underline{\mathbf{P}}_g^* - \underline{\mathbf{P}}_g) + \left[\underline{\mathbf{P}}_g\right]_\times \mathbf{R}\underline{\mathbf{u}}\sin\theta \\ \mathbf{R}\underline{\mathbf{u}}\sin\theta \end{bmatrix} \quad (16)$$

where $\underline{\mathbf{P}}_g$ is the center of gravity of the set of points, $\mathbf{R}$ is the rotation between a Cartesian frame defined by the points and the end-effector frame, and $\underline{\mathbf{u}}\theta$ are the axis and angle corresponding to the rotation matrix $\mathbf{R}^T\mathbf{R}^*$, that is the rotation between the current and desired orientation of the set of points.

In addition, the center of gravity of the set of points translates along a *straight line trajectory* from its initial to its final position in the camera frame. As a consequence, the features are most likely to remain in the camera field of view during the whole task.

4 Experimental results

The mobile manipulator of the Robotic Intelligence Lab consists of 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 miniature CMOS color cameras, linked to two video boards which deliver the visual features at video rate (30 Hz).

Figure 3: The stereo visual servoing manipulator setup.

The following table gives the estimation of the parameters (intrinsic and extrinsic) of both cameras, as used in the experiments.

F_u	F_v	b
300	450	118mm

The target object consists of four co-planar points located at the vertices of an 11cm square and the fifth point is located at the center of the square.

The velocity screw is computed from the pseudo-inverse of the jacobian matrix [4]:

$$\underline{\mathbf{v}} = -\lambda \mathbf{L}^+ (\underline{\mathbf{s}} - \underline{\mathbf{s}}^*) \quad (17)$$

with λ set to 0.5 in all the experiments.

Image measurements are noisy, since the experiments are carried out in a standard office environment, without any special illumination. As a result, there is an almost-uniform noise whose magnitude is ± 1 for u_i^l and u_i^r, and ± 2 for v_i^l and v_i^r. Additionally, pixel coordinates are quantified to a resolution of 200×200.

Experimental results are depicted in Figures 4. Each one consists of a set of plots (from top to bottom): the left image trajectories of the points, translation velocity, rotation velocity, and the 3D trajectory of the end-effector.

Convergence to the desired images is always achieved, but quality is worse with the stereo 2D features. As pointed out by Lamiroy et al. [8], the stereo jacobian is largely

overconstrained, and the control data $\underline{\mathbf{s}}$ and $\underline{\mathbf{s}}^*$ are redundant. But this is not sufficient to explain the curvy trajectory of the end-effector (bottom of Fig. 4), which almost leads out of the range of robot joints.

Such trajectory is neither caused by a too high gain: with $\lambda = 0.1$ a smoother but similar trajectory is obtained, as depicted in Fig. 4. This problem has not been addressed before since very few experiments with image-based stereo visual servoing have been carried out *with cameras mounted on the end-effector*. To our knowledge, only Maru et al. [9] have worked with this setup, but their tasks involved rather small rotations $(\phi, \theta, \psi) = (10, 10, 10)$ (degree). In our manipulation task, the rotation between the initial and destination poses is: $(\phi, \theta, \psi) = (72, 57, 50)$. Translational distance is 250 mm, as opposed to 173 mm in Maru *et al.* [9].

It is interesting to note that the trajectory with 2D and disparity is relatively close of the 3D trajectory. Approaches based on 3D features work better due to the linearity of the jacobian matrix. As shown theoretically, not only the image points but the center of gravity of 3D points translates along a straight line. As a result, the trajectory of the end-effector frame is closer to a straight line too, even with large rotations between frames. In summary, the use of 3D features allows the linearization of the jacobian and so a better joint decoupling.

5 Conclusion

In this paper, several approaches to image-based stereo visual servoing has been presented. Theoretical developments show how 3D control features are extracted from stereo images, and the jacobian matrix is computed for raw pixels, estimated 3D coordinates, and a new feature vector which uses stereo disparity.

As a main result, it has been shown how the effectiveness of the servoing task can be improved if estimated 3D features are used instead of raw image data. Real experiments with adverse conditions (large rotation, noisy images, coarse calibration) show that the trajectory of the end-effector strongly relies on the features chosen for the control loop.

Future work should state more precisely the robustness of the different approaches, with respect to camera parameters and signal loss. Furthermore, others visual features (i.e. lines) can be studying through the relationships between image data and estimated 3D features.

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] M. Asada, T. Tanaka, and K. Hosoda. Adaptive binocular visual servoing for independently moving target tracking. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 3, pages 2076 –2081, San Fransisco, USA, 2000. ICRA'00.

[2] E. Cervera, F. Berry, and P. Martinet. Stereo visual servoing with a single point: a comparative study. In *Proceedings of the IEEE International Conference on Advanced Robotics*, pages CD–ROM, Budapest, Hungary, 2001. ICAR'01.

[3] 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.

[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.

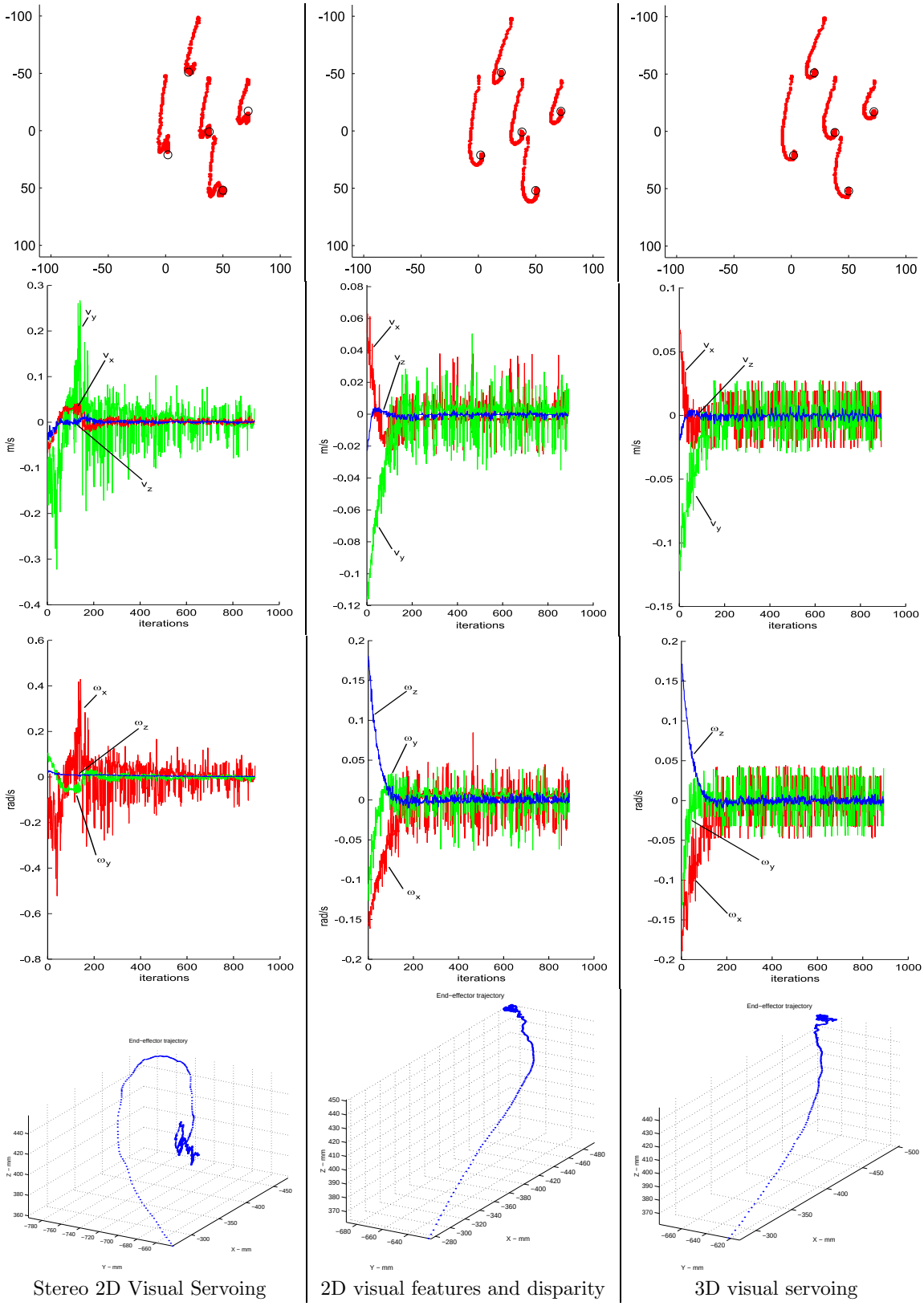
[5] E. Grosso, G. Metta, A. Oddera, and G. Sandini. Robust visual servoing in 3d reaching tasks. *IEEE Transactions on Robotics and Automation*, 12(5):732–742, october 1996.

[6] 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.

[7] S.H. Han, W.H. See, J. Lee, M.H. Lee, and H. Hashimoto. Image-based visual servoing control of a scara type dual-arm robot. In *Proceedings of the 2000 IEEE International Symposium on Industrial Electronics*, volume 2, pages 517 –522, Cholula, Puebla, Mexico, 2000. ISIE 2000.

[8] 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.

[9] 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.



Stereo 2D Visual Servoing | 2D visual features and disparity | 3D visual servoing

Figure 4: (Results from top to bottom)
Visual features trajectories in left image, Translation velocity,
Rotation velocity and Trajectory of the end-effector.