A Framework for Coupling Visual Control and Active Structure from Motion

Riccardo Spica, Paolo Robuffo Giordano, François Chaumette

A Framework for Coupling Visual Control and Active Structure from Motion

Riccardo Spica, Paolo Robuffo Giordano, and François Chaumette

I. SUMMARY

In most sensor-based robotic applications, the robot state can only be partially retrieved from onboard sensors and the use of estimation strategies is necessary for recovering online an approximation of any 'missing information' required to accurately control the robot action. With the exception of some trivial cases, however, the relationship between the sensor readings and the robot state is often nonlinear. As a consequence, and regardless of the particular estimation scheme, performance of the state estimation (e.g., convergence rate and/or final accuracy) depends, in general, on the particular trajectory followed by the sensor during the estimation process, with some trajectories being more informative than other ones.

The perspective projection performed by cameras is a classical example of such nonlinear sensor/state mapping. As well-known, a monocular camera cannot, e.g., estimate the depth of a point feature by traveling along the feature projection ray, and other constraints exist for different geometric primitives. This clearly creates a strong link between the motion performed by the robot/camera and the performance of any 3D structure estimation algorithm. Similarly, a poor accuracy in estimating the scene structure can also affect the performance of visual control schemes resulting in poor or even unstable closed-loop behaviors. Indeed, it has been shown in, e.g., [1] that a poor approximation of the 3D parameters of the scene can significantly affect stability of Image Based Visual Servoing (IBVS) controllers.

With respect to these considerations, in this contribution (which briefly summarizes [2]) we propose an *online* coupling between action and perception in the context of robot visual control by considering, in particular, the class of Image-Based Visual Servoing (IBVS) schemes [3] as representative case study. Indeed, besides being a widespread sensor-based technique, IBVS is also affected by *all* the aforementioned issues: on the one hand, whatever the chosen set of visual features (e.g., points, lines, planar patches), the associated *interaction matrix* always depends on some additional 3-D information not directly measurable from the visual input (e.g., the depth of a feature point). This 3-D information must then be approximated or estimated online via a Structure from Motion (SfM) algorithm, and an inaccurate knowledge (because of, e.g., wrong approximations or poor SfM performance) can degrade the servoing execution and also lead to instabilities or loss of feature tracking. On the other hand, the SfM performance is directly affected by the particular trajectory followed by the camera during the servoing [4]–[6]: the IBVS controller should then be able to realize the main visual task while, *at the same time*, ensuring a sufficient level of information gain for allowing an accurate state estimation.

In order to meet these objectives, we investigate a possible coupling between a recently developed framework for active SfM [5], [6] (the *active perception* component of our approach) and the execution of a standard IBVS task (the *visual control* component of our approach). The main idea is to project any optimization of the camera motion (aimed at improving the SfM performance) within the null-space of the considered task in order to not degrade the servoing execution. However, for any reasonable IBVS application, a simple null-space projection of a camera trajectory optimization turns out to be ineffective because of a structural lack of redundancy. Therefore, in order to gain the needed freedom for implementing the SfM optimization, we suitably exploit and extend the redundancy framework introduced in [7] which grants a *large* projection operator by considering the *norm* of the visual error as main task. In addition, we also propose an adaptive mechanism able to activate/deactivate online the camera trajectory optimization as a function of the accuracy of the estimated 3-D structure. Thanks to this addition, it is then possible to enable the SfM optimization only when strictly needed such as, e.g., when the 3-D estimation error grows larger than some desired minimum threshold.

In the following sections we give additional details on the active estimation (Sect. II) and redundancy resolution frameworks (Sect. III), followed by some experimental results (Sect. IV) of the proposed approach.

R. Spica is with the University of Rennes 1 at Irisa and Inria Rennes Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France riccardo.spica@irisa.fr

P. Robuffo Giordano is with the CNRS at Irisa and Inria Rennes Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France prg@irisa.fr

F. Chaumette is with Inria Rennes Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France francois.chaumette@irisa.fr

II. ACTIVE STRUCTURE FROM MOTION

In a general SfM estimation problem, a set of image measurements $\boldsymbol{s} \in \mathbb{R}^m$ (e.g. the normalized x and y coordinates of a tracked point feature on the image plane) is used to estimate some unmeasurable 3-D parameters $\boldsymbol{\chi} \in \mathbb{R}^p$ (e.g. the inverse depth for a point feature) exploiting knowledge of the camera linear and angular velocity $\boldsymbol{u} = [\boldsymbol{v}, \boldsymbol{\omega}]$ in the camera frame. The SfM system dynamics takes the general

form [6]:

$$\begin{cases} \dot{\boldsymbol{s}} = \boldsymbol{f}_m(\boldsymbol{s}, \boldsymbol{\omega}) + \boldsymbol{\Omega}^T(\boldsymbol{s}, \boldsymbol{v})\boldsymbol{\chi} \\ \dot{\boldsymbol{\chi}} = \boldsymbol{f}_u(\boldsymbol{s}, \boldsymbol{\chi}, \boldsymbol{u}) \end{cases} \tag{1}$$

where $\boldsymbol{f}_m \in \mathbb{R}^m$, $\boldsymbol{f}_u \in \mathbb{R}^p$ and $\boldsymbol{\Omega} \in \mathbb{R}^{p\times m}$ are known functions apart from the unknown value of $\boldsymbol{\chi}$ in $\boldsymbol{f}_u$. It has been shown in [6] that a nonlinear observer for (1) can be devised as:

$$\begin{cases} \dot{\hat{\boldsymbol{s}}} = \boldsymbol{f}_m(\boldsymbol{s}, \boldsymbol{\omega}) + \boldsymbol{\Omega}^T(\boldsymbol{s}, \boldsymbol{v})\hat{\boldsymbol{\chi}} + \boldsymbol{H}\boldsymbol{\xi} \\ \dot{\hat{\boldsymbol{\chi}}} = \boldsymbol{f}_u(\boldsymbol{s}, \hat{\boldsymbol{\chi}}, \boldsymbol{u}) + \alpha\boldsymbol{\Omega}(\boldsymbol{s}, \boldsymbol{v})\boldsymbol{\xi} \end{cases} \tag{2}$$

where $\boldsymbol{H}$ and α are free estimation gains and $\boldsymbol{\xi} = \boldsymbol{s} - \hat{\boldsymbol{s}}$ is the measurable part of the estimation error. By suitably tuning online the gain $\boldsymbol{H}$ one can assign a desired second-order transient behavior to the structure estimation error $\boldsymbol{z} = \boldsymbol{\chi} - \hat{\boldsymbol{\chi}}$. Moreover it can be shown that the convergence rate is determined by the eigenvalues of matrix $\alpha\boldsymbol{\Omega}(\boldsymbol{s}, \boldsymbol{v})\boldsymbol{\Omega}(\boldsymbol{s}, \boldsymbol{v})^T$. One can then increase *online* the SfM convergence by either (*i*) increasing the free gain α (at the cost of a possible higher level of noise), (*ii*) maximizing the eigenvalues σ_i^2 of $\boldsymbol{\Omega}(\boldsymbol{s}, \boldsymbol{v})\boldsymbol{\Omega}(\boldsymbol{s}, \boldsymbol{v})^T$ by *actively* acting on the camera linear velocity $\boldsymbol{v}$. The latter effect can be obtained using a gradient descent update of the velocity $\boldsymbol{v}$, i.e. by choosing a camera acceleration $\dot{\boldsymbol{v}} = k\nabla_{\boldsymbol{v}}\sigma_1^2$ where $\nabla_{\boldsymbol{v}}$ is the gradient with respect to $\boldsymbol{v}$ and $\sigma_1^2 \leq \sigma_2^2 \leq \cdots \leq \sigma_p^2$.

III. REDUNDANCY RESOLUTION

As outlined in Sect. II, the active estimation framework acts on the camera linear acceleration $\dot{\boldsymbol{v}}$. The classical first-order visual control framework described in, e.g, [3], must then be replaced by a second-order counterpart. Let us define $\boldsymbol{e} = \boldsymbol{s} - \boldsymbol{s}^*$ as the main visual task error, being $\boldsymbol{s}^*$ a constant desired value for the visual features $\boldsymbol{s} \in \mathbb{R}^m$. A second order control strategy for regulating $\boldsymbol{e}(t)$ to $\boldsymbol{0}$ exponentially while also optimizing a cost function $\mathcal{V}(\boldsymbol{u})$ is given by:

$$\dot{\boldsymbol{u}} = \dot{\boldsymbol{u}}_{\boldsymbol{e}} = \boldsymbol{L}^{\dagger}(-k_v\dot{\boldsymbol{e}} - k_p\boldsymbol{e} - \dot{\boldsymbol{L}}\boldsymbol{u}) + (\boldsymbol{I}_6 - \boldsymbol{L}^{\dagger}\boldsymbol{L})\boldsymbol{r} \tag{3}$$

where $k_p > 0$, $k_v > 0$, $\boldsymbol{L}(\boldsymbol{s}, \hat{\boldsymbol{\chi}}) \in \mathbb{R}^{m\times 6}$ is the feature interaction matrix ($\dot{\boldsymbol{s}} = \boldsymbol{L}\boldsymbol{u}$) and $\boldsymbol{r} = k\nabla_{\boldsymbol{u}}\mathcal{V}(\boldsymbol{u})$. Note that the unmeasurable 3-D quantities $\boldsymbol{\chi}$ must be replaced by their estimation provided, e.g., by the observer (2). By setting, e.g., $\mathcal{V}(\boldsymbol{u}) = \sigma_1^2(\boldsymbol{v}, \boldsymbol{s})$ one would obtain an optimization of the estimation convergence rate to the extent that is compatible with the main visual task. In most visual servoing applications, however, the visual task is *overconstraining*, i.e. $m > 6$ and, in general, $\mathrm{rank}\,(\boldsymbol{L}) = 6$ implying $(\boldsymbol{I}_6 - \boldsymbol{L}^{\dagger}\boldsymbol{L}) = 0$ and thus the secondary task will never be realized.

A workaround able to 'increase' the system redundancy is to consider, as suggested in [7], the regulation of the (scalar) norm of the visual error $\nu = \|\boldsymbol{e}\|$ as a main task instead of $\boldsymbol{e}$. It is easy to show that $\dot{\nu} = \frac{1}{\nu}\boldsymbol{e}^T\dot{\boldsymbol{e}}$ and hence a second order control strategy for regulating $\nu(t)$ to $\boldsymbol{0}$ exponentially while also optimizing a cost function $\mathcal{V}(\boldsymbol{u})$ can be devised as:

$$\dot{\boldsymbol{u}} = \dot{\boldsymbol{u}}_{\nu} = \boldsymbol{J}_{\nu}^{\dagger}(-k_v\dot{\nu} - k_p\nu - \dot{\boldsymbol{J}}_{\nu}\boldsymbol{u}) + (\boldsymbol{I}_6 - \boldsymbol{J}_{\nu}{}^{\dagger}\boldsymbol{J}_{\nu})\boldsymbol{r} \tag{4}$$

where $k_p > 0$, $k_v > 0$, $\boldsymbol{J}_{\nu} = \frac{1}{\nu}\boldsymbol{e}^T\boldsymbol{L} \in \mathbb{R}^{1\times 6}$. In this case then one has $\mathrm{rank}\,(\boldsymbol{I}_6 - \boldsymbol{J}_{\nu}{}^{\dagger}\boldsymbol{J}_{\nu}) \geq 5$ and the redundant degrees of freedom can be used to optimize the estimation convergence rate through the action of $\boldsymbol{r}$ (i.e., to select *online* a camera trajectory which is more informative about the 3D estimation task). A shortcoming of this strategy is that $\boldsymbol{J}_{\nu}$ becomes singular when $\nu \approx 0$ and one has to switch to (3) when close to convergence. As shown in [8], care has to be taken when switching from (4) to (3): an intermediate control phase must be introduced to guarantee that $\boldsymbol{e}$ and $\dot{\boldsymbol{e}}$ are aligned before the switch so that the error norm ν keeps converging monotonically to zero.

IV. EXPERIMENTAL RESULTS

As representative case study we consider the regulation of $N = 4$ point features with, thus, $\boldsymbol{s} = (x_1, y_1, \ldots, y_N) \in \mathbb{R}^m$, and $\boldsymbol{L}_{\boldsymbol{s}} = (\boldsymbol{L}_{\boldsymbol{s}_1}, \ldots, \boldsymbol{L}_{\boldsymbol{s}_N}) \in \mathbb{R}^{m\times 6}$, $m = 8 > 6$ (overconstraining task), with $\boldsymbol{L}_{\boldsymbol{s}_i}$ being the standard 2×6 interaction matrix for a point feature [3]. As for vector $\boldsymbol{\chi}$, we then have $\boldsymbol{\chi} = (\chi_1, \ldots, \chi_N) \in \mathbb{R}^p$, $p = 4$, where $\chi_i = 1/Z_i$ as explained in Sect. II. To demonstrate the effectiveness of our approach we show in Fig. 1 the results obtained in the following four different experimental cases, all starting from the same initial conditions:

1 (blue lines): the control law (4) is implemented until close to convergence when a switch to (3) is necessary. The estimator (2) runs in parallel to the servoing task for generating $\hat{\boldsymbol{\chi}}(t)$, fed to all the various control terms. The active optimization of the camera motion takes place while (4) is used by setting $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2$ and

$$\mathcal{V}(\boldsymbol{v}) = k_\sigma\gamma\log\left(\frac{\gamma+\sigma^2(\boldsymbol{v})}{\gamma}\right) - \frac{k_d}{2}\|\boldsymbol{v}\|^2, \quad \gamma, k_d, k_\sigma > 0;$$

2 (red lines): the classical control law (3) is implemented. The estimator (2) is *still* run in parallel to the servoing task for generating $\hat{\boldsymbol{\chi}}(t)$, but no optimization of the estimation error convergence is performed;

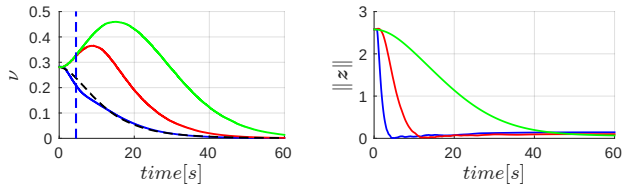
3 (green lines): the classical control law (3) is again implemented, however, the estimator is *not* run and vector $\hat{\boldsymbol{\chi}}(t)$ is taken coincident with its value at the desired pose, i.e., $\hat{\boldsymbol{\chi}}(t) = \boldsymbol{\chi}^* = const$, as customary in many visual servoing applications;

4 (black lines): the classical control law (3) is again implemented but by exploiting knowledge of the ground truth value $\hat{\boldsymbol{\chi}}(t) = \boldsymbol{\chi}(t)$ during the whole servoing execution. This case represents a reference ideal 'ground truth'.

Figure 1 clearly shows on one hand the beneficial effects of a good approximation of $\boldsymbol{\chi}$ on the visual task performance and, on the other hand the additional improvement resulting from taking active measures aiming at optimizing the camera motion to maximize the estimation performance. A video of these (and additional) experimental results is available at https://youtu.be/BhwoaNTfpvc.

REFERENCES

[1] E. Malis, Y. Mezouar, and P. Rives, "Robustness of Image-Based Visual Servoing With a Calibrated Camera in the Presence of Uncertainties in the Three-Dimensional Structure," *IEEE Trans. on Robotics*, vol. 26, no. 1, pp. 112–120, Feb 2010.



(a) (b)

Fig. 1: Experimental results. (a): error norm $\nu(t)$ in the four experimental cases. The vertical dashed blue line represents the switch from (4) to (3). (b): approximation error $\|\boldsymbol{z}(t)\| = \|\boldsymbol{\chi}(t) - \hat{\boldsymbol{\chi}}(t)\|$.

[2] R. Spica, P. Robuffo Giordano, and F. Chaumette, "Bridging Visual Control and Active Perception," *in preparation for IEEE Trans. on Robotics*, 2015.

[3] F. Chaumette and S. Hutchinson, "Visual servo control, Part I: Basic approaches," *IEEE Robotics and Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.

[4] A. Martinelli, "Vision and IMU data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination," *IEEE Trans. on Robotics*, vol. 1, no. 28, pp. 44–60, 2012.

[5] R. Spica and P. Robuffo Giordano, "A Framework for Active Estimation: Application to Structure from Motion," in *52nd IEEE Conf. on Decision and Control*, 2013, pp. 7647–7653.

[6] R. Spica, P. Robuffo Giordano, and F. Chaumette, "Active Structure from Motion: Application to Point, Sphere and Cylinder," *IEEE Trans. on Robotics*, vol. 30, no. 6, pp. 1499–1513, 2014.

[7] M. Marey and F. Chaumette, "A new large projection operator for the redundancy framework," in *2010 IEEE Int. Conf. on Robotics and Automation*, 2010, pp. 3727–3732.

[8] R. Spica, P. Robuffo Giordano, and F. Chaumette, "Coupling Visual Servoing with Active Structure from Motion," in *2014 IEEE Int. Conf. on Robotics and Automation*, Hong Kong, China, May 2014, pp. 3090–3095.