Vision-based Manipulation of Deformable and Rigid Objects Using Subspace Projections of 2D Contours

Jihong Zhu, David Navarro-Alarcon, Robin Passama, Andrea Cherubini

Vision-based Manipulation of Deformable and Rigid Objects Using Subspace Projections of 2D Contours

Jihong Zhu, David Navarro-Alarcon, Robin Passama and Andrea Cherubini

***Abstract*—This paper proposes a unified vision-based manipulation framework using image contours of deformable/rigid objects. Instead of using human-defined cues, the robot automatically learns the features from processed vision data. Our method simultaneously generates—from the same data—both, visual features and the interaction matrix that relates them to the robot control inputs. Extraction of the feature vector and control commands is done online and adaptively, with little data for initialization. The method allows the robot to manipulate an object without knowing whether it is rigid or deformable. To validate our approach, we conduct numerical simulations and experiments with both deformable and rigid objects.**

***Index Terms*—Visual servoing, manipulation, deformable objects, sensor-based control.**

I. Introduction

HUMANS are capable of manipulating both rigid and deformable objects. However, robotic researchers tend to consider the manipulation of these two classes of objects as separate problems. This paper presents our efforts in formulating a generalized framework for vision-based manipulation of both rigid and deformable objects, which does not require prior knowledge of the object's mechanical properties.

In classical visual servoing literature [1], the vector $\boldsymbol{s}$ denotes the set of features selected to represent the object in the image. These features represent both the object's pose and its shape. We denote the process of selecting $\boldsymbol{s}$ as *parameterization*. The aim of visual servoing is to minimize, through robot motion, the feedback error $\boldsymbol{e} = \boldsymbol{s}^* - \boldsymbol{s}$ between the desired $\boldsymbol{s}^*$ and the current (i.e., measured) feature $\boldsymbol{s}$.

One of the initial works on vision-based manipulation of deformable objects is presented in [2] to solve a knotting problem by a topological model. Smith et al. developed a relative elasticity model, such that vision can be utilized without a physical model for the manipulation task [3]. A classical model-free approach in manipulating deformable objects is developed in [4]. More recent research [5] proposes a method for online estimation of the deformation Jacobian, based on weighted least square minimization with a sliding window. In [6] and [7], the vision-based deformable objects manipulation is termed as *shape servoing*. An expository paper on the topic is available in [8]. A recent work on shape servoing of plastic material was presented in [9].

For *shape servoing*, commonly selected features are curvatures [6], points [10] and angles [11]. Laranjeira et al. proposed a catenary-based feature for tethered management on wheeled and underwater robots [12], [13]. A more general feature vector is that containing the Fourier coefficients of the object contour [7] and [14]. Yet, all these approaches require the user to specify a model, e.g., the object geometry [6], [10], [11] or a function [7], [12], [14] for selecting the feature. Alternative data-driven (hence, modeling-free) approaches rely on machine learning. Nair et al. combine learning and visual feedback to manipulate ropes in [15]. The authors of [16] employ deep neural networks to manipulate deformable objects given their 3D point cloud. All these methods rely on (deep) connectionists models that invariably require training through an extensive data set. The collected data has to be diverse enough to generalize the model learnt by this type of networks. Furthermore, the above-mentioned methods focus only on deformable object manipulation.

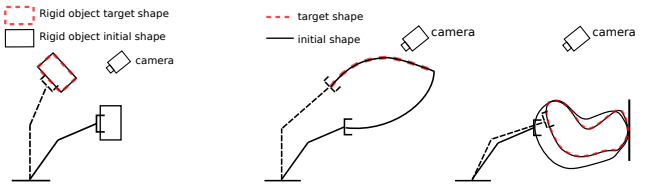
As for *pose control of rigid objects*, the trend in visual servoing is to find features which are independent from the object characteristics. Following this trend, [17] proposes the use of image moments. More recently, researchers have proposed direct visual servoing (DVS) methods, which eliminate the need for user-defined features and for the related image processing procedures. The pioneer DVS works [18], [19] propose using the whole image luminance to control the robot, leading to "photometric" visual servoing. Bakthavatchalam et al. join the two ideas by introducing photometric moments [20]. A subspace method [21] can further enhance the convergence of photometric visual servoing, via Principal Component Analysis (PCA). This method was first introduced for visual servoing in [22]. In that work, using an eye-in-hand setup, the image was compressed to obtain a low-dimensional vector for controlling the robot to a desired pose. Similarly, the authors of [23] transformed the image into a lower dimensional hyper surface, to control the robot position via in-hand camera feedback. However, DVS generally considers rigid and static scenes, where the robot controls the motion of the camera (eye-in-hand setup) to change only the image viewpoint, and not the environment. This setup avoids breaking the Lambertian hypotheses, which are needed for the methods to be applicable. For this reason, to our knowledge, DVS was never applied to object manipulation, since changes in the pose and/or shape of the object would break the Lambertian assumption.

Compared with above-mentioned works, our paper has the following original contributions:

This work is supported in part by the EU H2020 research and innovation programme as part of the project VERSATILE under grant agreement No 731330, by the Research Grants Council (RGC) of Hong Kong under grant number 14203917, and by the PROCORE-France/Hong Kong RGC Joint Research Scheme under grant F-PolyU503/18.

J. Zhu, R. Passama and A. Cherubini are with LIRMM - Université de Montpellier CNRS, 161 Rue Ada, 34090 Montpellier, France. firstname.lastname@lirmm.fr

D. Navarro-Alarcon is with The Hong Kong Polytechnic University, Department of Mechanical Engineering, Kowloon, Hong Kong. dna@ieee.org



(a) Rigid objects (b) Deformable objects

Fig. 1. Vision-based manipulation of rigid and deformable objects. For rigid objects (left): control pose (translation and rotation). For deformable objects (right): control the pose, and also shape.

1) We introduce a new compact feedback feature vector – based on PCA of sampled 2D contours – which can be used for both deformable and rigid object manipulation.
2) We exploit the linear properties of PCA and of the local interaction matrix, to initialize our algorithm with little data – the same data for feature vector extraction and for interaction matrix estimation.
3) We report experiments using the same framework to manipulate objects with different unknown geometric and mechanical properties.

In a nutshell, we propose an efficient data-driven approach that allows manipulation of an object regardless of its deformation characteristics. To the best of authors' knowledge, there has not been any similar framework proposed before.

The paper is organized as follows. Sect. II presents the problem. Sect. III outlines the framework. Sect. IV elaborates on the methods. In Sect. V, we analyze and verify the methods by numerical simulations. Then, Sect. VI presents the robotic experiments and we conclude in Sect. VII.

II. Problem Statement

In this work, we aim at solving object manipulation tasks with visual feedback. We rely on the following hypotheses:

- The shape and pose of the object are represented by its 2-D contour on the image as seen from a camera fixed in the robot workspace (eye-to-hand setup). We denote this contour as
$$\boldsymbol{c} = [\boldsymbol{p}_1 \ \cdots \ \boldsymbol{p}_K]^T \in \mathbb{R}^{2K}, \tag{1}$$
where $\boldsymbol{p}_j = [u_j \ v_j] \in \mathbb{I}$ denotes the jth pixel of the contour in the image $\mathbb{I}$.
- The contour is always entirely visible in the scene and there are no occlusions.
- One of the robot's end-effectors holds one point of the object (we consider the grasping problem to be already solved). At each control iteration i, its pose is $\boldsymbol{r}_i \in \mathbb{SE}(3)$, and it can execute motion commands $\delta\boldsymbol{r}_i \in \mathbb{SE}(3)$ that drive the robot as $\boldsymbol{r}_{i+1} = \boldsymbol{r}_i + \delta\boldsymbol{r}_i$.

Problem Statement. Given a desired shape of the object, represented by its contour $\boldsymbol{c}^*$, we aim at designing a controller that generates a sequence of robot motions $\delta\boldsymbol{r}$ to drive the initial contour to the desired one, by relying on visual feedback.

The controller should work without any knowledge of the object physical characteristics (i.e., for both rigid and deformable objects). Typically, since rigid and deformable objects behave differently during manipulation, we set the following manipulation goals:

- Rigid objects: move them to a desired pose (see Fig. 1a).
- Deformable objects: move them to a desired pose with a desired shape (see Fig. 1b).

III. Framework Overview

In this section, we present an overview of the proposed approach motivated by the problem analysis.

A. Problem analysis

We can work directly on the object shape space by selecting the contour as the feature vector $\boldsymbol{s} \equiv \boldsymbol{c} \in \mathbb{R}^{2K}$. However, this will result in an unnecessarily large dimension of the feature vector (e.g., if $K = 50$, $\boldsymbol{s}$ has 100 components). The high dimensional feature vector increases the computation demand and complicates the control due to the high under-actuation of the system. Therefore, instead of working directly on shape space, we work on a feature vector space that has reduced dimensions. For that, we split the problem into two sub-problems: *parameterization* and *control*, see Fig. 2.

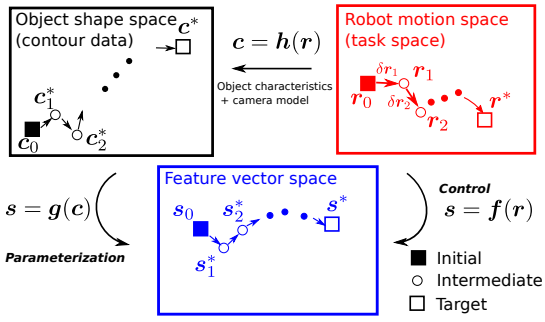


Fig. 2. Graphic representation of the vision-based manipulation problem, with its two sub-problems, *parameterization* and *control*.

Parameterization consists in representing the shape contour via a compact feature vector $\boldsymbol{s} \in \mathbb{R}^k$, such that $k << 2K$. We denote this representation as $\boldsymbol{s} = \boldsymbol{g}(\boldsymbol{c})$.

Control consists in computing robot motions $\delta\boldsymbol{r}_1, \delta\boldsymbol{r}_2, \ldots$, so that the object's representation $\boldsymbol{s}$ converges to a desired target $\boldsymbol{s}^*$. *Control* can be broken down to solving the optimization problem:
$$\boldsymbol{r}^* = \arg\min_{\boldsymbol{r}} (\boldsymbol{f}(\boldsymbol{r}) - \boldsymbol{s}^*) \tag{2}$$
where $\boldsymbol{s} = \boldsymbol{f}(\boldsymbol{r})$ denotes the mapping between robot pose and feature vector, which is assumed to be smooth and often nonlinear. If we know the analytic solution to $\boldsymbol{f}(\boldsymbol{r})$, we can solve (2) and obtain the target shape in a single iteration by commanding $\boldsymbol{r}^*$. However, the full mapping $\boldsymbol{f}(\boldsymbol{r})$ is the result of two subsequent mappings. A mapping $\boldsymbol{c} = \boldsymbol{h}(\boldsymbol{r})$ exists between the robot pose $\boldsymbol{r}$ and the resulting contour $\boldsymbol{c}$. This, combined with the *parameterization* mapping above yields: $\boldsymbol{f}(\boldsymbol{r}) = \boldsymbol{g}(\boldsymbol{h}(\boldsymbol{r}))$. To solve for $\boldsymbol{f}(\boldsymbol{r})$, one needs to have the analytic expressions of both $\boldsymbol{h}$ and $\boldsymbol{g}$. While the latter comes from the *parameterization*, the former ($\boldsymbol{h}$)

is difficult to obtain, since it encompasses both the object characteristics and the camera projection model. For different objects and camera poses, we expect different properties and camera parameters, therefore a different $\boldsymbol{h}$. Even for $\boldsymbol{g}$, we would like our framework to automatically extract the feature vector without explicit human definition. Therefore, unlike traditional approaches where $\boldsymbol{g}$ is user-defined and known, in our framework, the robot has to infer $\boldsymbol{g}$ from sensor data, which in return makes deriving $\boldsymbol{g}$ difficult.

A solution to this problem is to approximate the full mapping $\boldsymbol{f}(\boldsymbol{r})$ from sensor observations. Classic data-driven approaches typically require a long training phase to collect vast and diverse data for approximating $\boldsymbol{f}(\boldsymbol{r})$. In some cases (robotics surgery for instance), it is not possible to collect such data beforehand. Moreover, if the object changes, new data has to be collected to retrain the model, leading to a cumbersome process. In this paper instead, we aim at doing the data collection online, with minimum initialization.

Thus, instead of estimating the full nonlinear mapping $\boldsymbol{f}(\boldsymbol{r})$, we divide it into piece-wise linear models [24] at successive equilibrium points. These models are considered time invariant in the neighbourhood of the equilibrium points. We then compute the control law for each linear model and apply it to the robot end-effector.

B. Proposed methods

Given a target shape $\boldsymbol{c}^*$, we define an intermediate local target $\boldsymbol{c}_i^*$ at each $i = 1, 2, \dots$ (see Fig. 2). At the i^{th} instance, the robot autonomously generates a local mapping $\boldsymbol{g}_i$ to produce the feature vector $\boldsymbol{s}_i = \boldsymbol{g}_i(\boldsymbol{c}_i)$. The robot then finds the local mapping $\boldsymbol{s}_i = \boldsymbol{f}_i(\boldsymbol{r}_i)$ online.

Consider at the current time instant i, the shape $\boldsymbol{c}_i$, the intermediate target $\boldsymbol{c}_i^*$ and the local parameterization $\boldsymbol{g}_i$, we can transform shape data into a feature vector by:

$$\boldsymbol{s}_i = \boldsymbol{g}_i(\boldsymbol{c}_i), \;\; \boldsymbol{s}_i^* = \boldsymbol{g}_i(\boldsymbol{c}_i^*). \tag{3}$$

The linearized version of $\boldsymbol{s} = \boldsymbol{f}(\boldsymbol{r})$ centered at $(\boldsymbol{s}_i, \boldsymbol{r}_i)$ is then:

$$\delta \boldsymbol{s}_i = \boldsymbol{L}_i \delta \boldsymbol{r}_i, \tag{4}$$

with

$$\begin{aligned} \boldsymbol{L}_i &= \frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{r}} \,|_{\boldsymbol{r}=\boldsymbol{r}_i}, \\ \delta \boldsymbol{s}_i &= \boldsymbol{s}_{i+1} - \boldsymbol{s}_i, \\ \delta \boldsymbol{r}_i &= \boldsymbol{r}_{i+1} - \boldsymbol{r}_i. \end{aligned} \tag{5}$$

The matrix $\boldsymbol{L}_i$ represents a local mapping, referred to as the interaction matrix in the visual servoing literature [1]. If $\boldsymbol{L}_i$ can be estimated online at each iteration i, then, we can design one-step control laws to drive $\boldsymbol{s}_i$ towards $\boldsymbol{s}_i^*$.

After the robot has executed the motion command $\delta \boldsymbol{r}_i$, we update the next target to be $\boldsymbol{s}_{i+1}^*$, and so on, until it reaches the final (desired) target $\boldsymbol{s}^*$. Although the validity region of this local linear mapping is much smaller than that of the original nonlinear mapping, it results in an online training with less data and reduced computational demand.

In the following section, we detail our proposed approach.

IV. Methods

Figure 3 shows the building blocks for the overall framework. In this section, we focus on the red dashed part of the diagram. We will elaborate on each red block in the subsequent subsections. The blue block represents the image processing pipeline that will be discussed in Sect. VI-A.

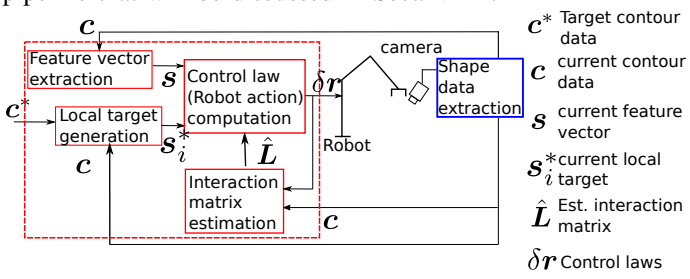


Fig. 3. The block diagram that represents the overall framework.

A. Feature vector extraction

There are many ways for parameterizing $\boldsymbol{c}$ to reduce its dimension. Since the interaction matrix in (4) represents a *linear* mapping for the control, to be consistent, we look for a *linear* transformation for the parameterization.

One of the prominent linear dimension reduction methods is Principal Component Analysis (PCA). PCA finds a new orthogonal basis for high-dimensional data. This enables projection of the data to lower dimension with the minimal sum of squared residuals. We apply PCA to map $\boldsymbol{c} \in \mathbb{R}^{2K}$ to $\boldsymbol{s} \in \mathbb{R}^k$.

To find the projection, we collect M images with different shapes of the object and construct the data matrix $\boldsymbol{\Gamma} = [\boldsymbol{c}_1 \;\; \boldsymbol{c}_2 \;\; \cdots \;\; \boldsymbol{c}_M] \in \mathbb{R}^{2K \times M}$. Then, we shift the columns of $\boldsymbol{\Gamma}$ by the mean contour $\bar{\boldsymbol{c}} = \sum_{i=1}^{M} \boldsymbol{c}_i / M$:

$$\bar{\boldsymbol{\Gamma}} = [\boldsymbol{c}_1 - \bar{\boldsymbol{c}} \quad \boldsymbol{c}_2 - \bar{\boldsymbol{c}} \quad \cdots \; \boldsymbol{c}_M - \bar{\boldsymbol{c}}] \in \mathbb{R}^{2K \times M}. \tag{6}$$

We then compute the covariance matrix $\boldsymbol{C} = \bar{\boldsymbol{\Gamma}}^T \bar{\boldsymbol{\Gamma}}$, and apply Singular Value Decomposition (SVD) to it:

$$\boldsymbol{C} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^T. \tag{7}$$

Once we have obtained the eigenvector matrix $\boldsymbol{U} \in \mathbb{R}^{2K \times 2K}$, we can move on to select the first k columns of $\boldsymbol{U}$ denoted by $\boldsymbol{U}(k) \in \mathbb{R}^{2K \times k}$. Then, the $2K$-dimensional contour $\boldsymbol{c}$ can be projected into a smaller k-dimensional feature vector $\boldsymbol{s}$ as:

$$\boldsymbol{s} = \boldsymbol{U}^T(k)(\boldsymbol{c} - \bar{\boldsymbol{c}}) \in \mathbb{R}^k. \tag{8}$$

To assess the quality of this projection, we can compute the *explained variance* using the eigenvalue matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{2K \times 2K}$ in (7). Denoting the diagonal entries of $\boldsymbol{\Sigma}$ as $\sigma_1, \cdots, \sigma_{2K}$, the explained variance of the first k components is:

$$\Upsilon(k) = \frac{\sum_{j=1}^{k} \sigma_j}{\sum_{j=1}^{2K} \sigma_j}. \tag{9}$$

where Υ is a scalar between 0 and 1 (since $\sigma_j > 0, \; \forall j$), indicating to what extent the k components represent the original data (a larger Υ suggests a better representation).

At this stage, it should be clear for the reader that there is a crucial trade-off between choosing a low or high value for the number of features, k. A low k will ease controllability (given the limited robot inputs), whereas a high k will improve the data representation (by increasing Υ). In the next section, we explain how to select the best value of k.

B. Selection of the Feature Dimension k

Feature vector $\boldsymbol{s} \in \mathbb{R}^k$ in (8) lies in the new orthonormal basis represented by the columns of $\boldsymbol{V}$ in (7). Therefore, both $\boldsymbol{s}$ and its variation $\delta\boldsymbol{s}$ span a space of dimension k. Similarly, $\boldsymbol{r}$ and $\delta\boldsymbol{r}$ span a space of dimension n.

From (4), we know that at each iteration, $\delta\boldsymbol{s} \in \mathbb{R}^k$ is the result of a *constant* linear mapping ($\boldsymbol{L}$) on $\delta\boldsymbol{r} \in \mathbb{R}^n$. For an arbitrary number p of samples of small robot motions $\delta\boldsymbol{r}$: $\boldsymbol{R} = \begin{bmatrix}\delta\boldsymbol{r}_1 \cdots \delta\boldsymbol{r}_p\end{bmatrix}$ such that the mapping $\boldsymbol{L}$ stays constant, we have $\text{rank}(\boldsymbol{R}) \leq n$. Using (4) we obtain

$$\boldsymbol{S} = \boldsymbol{L}\boldsymbol{R} \tag{10}$$

such that $\boldsymbol{S} = \begin{bmatrix}\delta\boldsymbol{s}_1 \cdots \delta\boldsymbol{s}_p.\end{bmatrix}$. Each column of $\boldsymbol{S}$ is the result of corresponding robot motion. To find the relationship between k and n, we introduce the following lemma [25]:

Lemma 1. $\text{rank}(\boldsymbol{A}\boldsymbol{B}) \leq \min(\text{rank}(\boldsymbol{A}), \text{rank}(\boldsymbol{B}))$.

Applying *Lemma 1* to (10):

$$\begin{aligned}\text{rank}(\boldsymbol{S}) &\leq \min(\text{rank}(\boldsymbol{L}), \text{rank}(\boldsymbol{R})) \\ &\leq \text{rank}(\boldsymbol{R}) \leq n.\end{aligned} \tag{11}$$

Since $\delta\boldsymbol{s} \in \mathbb{R}^k$, the maximum dimension of $\text{rank}(\boldsymbol{S})$ is k: $\text{rank}(\boldsymbol{S}) \leq k$. According to (11), $k \leq n$. Since a larger k yields a better approximation of the real shape data, the value of k should be chosen as large as possible considering the condition $k \leq n$. Therefore we choose $k = n$.

C. Local Target Generation

Let us now explain how we generate a local desired contour $\boldsymbol{c}_i^*$ given current contour $\boldsymbol{c}_i$ and final desired contour $\boldsymbol{c}^*$ (Algorithm 1). The overall shape error is given by:

$$\boldsymbol{c_e} = \boldsymbol{c}^* - \boldsymbol{c}_i. \tag{12}$$

We define the intermediate desired contour as:

$$\boldsymbol{c}_i^* = \boldsymbol{c}_i + \frac{1}{\eta}\boldsymbol{c_e}, \tag{13}$$

with $\eta = 1, 2, \ldots$ an integer that ensures that $\boldsymbol{c}_i^*$ is a "good" local target for $\boldsymbol{c}_i$ (i.e., the two are similar). This means that if we project the intermediate local data using $\boldsymbol{U}_i \in \mathbb{R}^{2K\times 2K}$ (note that we are using the full projection matrix and not just the first k columns), the feature vector $\boldsymbol{s}_i^* = \boldsymbol{U}_i(\boldsymbol{c}_i^* - \bar{\boldsymbol{c}}) \in \mathbb{R}^{2K}$ should fulfil:

$$\Psi(k) = \frac{\sum_{j=1}^{k}\left|s_{i,j}^*\right|}{\sum_{j=1}^{2K}\left|s_{i,j}^*\right|} \geq \epsilon, \tag{14}$$

with $\epsilon \in]0;1[$ a threshold and $s_{i,j}^*$ the j-th component of $\boldsymbol{s}_i^*$.

Algorithm 1 outlines the steps for computing the local intermediate targets, so that:

- they are near the final target,
- the corresponding feature vector can be extracted with the current learned projection matrix.

Remark 1. The reachability of a local target can only be verified with a global deformation model which we want to avoid identifying in our methods. We will further discuss this issue in the Conclusion (Sect. VII).

Algorithm 1 Local target generation

```
localTargetFound = false
Ψ_0 = 0
η = 1
while not localTargetFound do
    c_i^* = c_i + (1/η)(c^* − c_i)
    s_i^* = U_i c_i^*
    Ψ_η = Σ_{j=1}^{k} |s_{i,j}^*| / Σ_{j=1}^{2K} |s_{i,j}^*|
    if Ψ_η ≥ ε or Ψ_η < Ψ_{η−1} then
        localTargetFound = true
    end if
    η = η + 1
end while
```

D. Interaction Matrix Estimation

Let us consider the current contour $\boldsymbol{c}_i$ and the local target $\boldsymbol{c}_i^*$. In this section, we show how we can implement the PCA and model estimation together and online. We denote the robot motions and corresponding object contours over the last M iterations (prior to iteration $i \geq M$) as:

$$\begin{aligned}\boldsymbol{\Delta R}_i &= \begin{bmatrix}\delta\boldsymbol{r}_{i-M+1}\ \delta\boldsymbol{r}_{i-M+2} \cdots \delta\boldsymbol{r}_i\end{bmatrix} \in \mathbb{R}^{n\times M} \\ \boldsymbol{\Gamma}_i &= \begin{bmatrix}\boldsymbol{c}_{i-M}\ \boldsymbol{c}_{i-M+1}\ \boldsymbol{c}_{i-M+2} \cdots \boldsymbol{c}_i\end{bmatrix} \in \mathbb{R}^{2K\times(M+1)}\end{aligned} \tag{15}$$

By selecting $k = n$, we compute the projection matrix $\boldsymbol{U}_i(n) \in \mathbb{R}^{2K\times n}$, from $\boldsymbol{\Gamma}_i$ and $\bar{\boldsymbol{c}}_i$ via (6) and (7). Then, using $\boldsymbol{U}_i$, we project current contour $\boldsymbol{c}_i$, desired contour $\boldsymbol{c}_i^*$ and shape matrix $\boldsymbol{\Gamma}_i$:

$$\begin{aligned}&\boldsymbol{s}_i = \boldsymbol{U}_i(n)^T(\boldsymbol{c}_i - \bar{\boldsymbol{c}}_i) \in \mathbb{R}^n, \\ &\boldsymbol{s}_i^* = \boldsymbol{U}_i(n)^T(\boldsymbol{c}_i^* - \bar{\boldsymbol{c}}_i) \in \mathbb{R}^n, \\ &\boldsymbol{S}_i = \boldsymbol{U}_i(n)^T\bar{\boldsymbol{\Gamma}}_i = \begin{bmatrix}\boldsymbol{s}_{i-M}\ \boldsymbol{s}_{i-M+1}\ \cdots\ \boldsymbol{s}_i\end{bmatrix} \in \mathbb{R}^{n\times(M+1)}.\end{aligned} \tag{16}$$

In (16), $\bar{\boldsymbol{\Gamma}}_i$ is normalized by $\bar{\boldsymbol{c}}_i$ as in (6). We can then compute $\boldsymbol{\Delta S}_i$ from (16), by subtracting consecutive columns of $\boldsymbol{S}_i$:

$$\boldsymbol{\Delta S}_i = \begin{bmatrix}\delta\boldsymbol{s}_{i-M+1}\ \delta\boldsymbol{s}_{i-M+2} \cdots \delta\boldsymbol{s}_i\end{bmatrix} \in \mathbb{R}^{n\times M}. \tag{17}$$

Using $\boldsymbol{\Delta S}_i \in \mathbb{R}^{n\times M}$ and $\boldsymbol{\Delta R}_i \in \mathbb{R}^{n\times M}$ we can now estimate the local interaction matrix $\boldsymbol{L}_i \in \mathbb{R}^{n\times n}$ at iteration i. We assume that near this iteration, the system remains linear and time invariant: $\boldsymbol{L}_i$ is constant. Using the local linear model (4), we can write the following:

$$\boldsymbol{\Delta S}_i = \boldsymbol{L}_i\boldsymbol{\Delta R}_i. \tag{18}$$

Our goal then is to solve for $\boldsymbol{L}_i$, given $\boldsymbol{\Delta S}_i$ and $\boldsymbol{\Delta R}_i$. Note that this is an overdetermined linear system (with $n \times M$ equations for n^2 unknonwns). Let us consider $\boldsymbol{\Delta R}_i \in \mathbb{R}^{n\times M}$ has full row rank. Note this sufficiently implies $M \geq n$. With this prerequisite, $\text{rank}(\boldsymbol{\Delta R}_i) = n$. Therefore, $\text{rank}(\boldsymbol{\Delta R}_i\boldsymbol{\Delta R}_i^T) = n$, and its inverse exists. We post multiply (18) by $\boldsymbol{R}_i^T$:

$$\boldsymbol{\Delta S}_i\boldsymbol{R}_i^T = \boldsymbol{L}_i\boldsymbol{\Delta R}_i\boldsymbol{\Delta R}_i^T. \tag{19}$$

Then, since $\boldsymbol{\Delta R}_i\boldsymbol{\Delta R}_i^T$ is invertible, the $\boldsymbol{L}_i$ that best fulfills (18) is:

$$\hat{\boldsymbol{L}}_i = \boldsymbol{\Delta S}_i\boldsymbol{\Delta R}_i^T(\boldsymbol{\Delta R}_i\boldsymbol{\Delta R}_i^T)^{-1}. \tag{20}$$

This matrix is also the solution of the optimization problem

$$\min_{\boldsymbol{L}_i} \sum_{j=1}^{n} \left\| \delta\boldsymbol{\sigma}_j^T - \Delta\boldsymbol{R}_i^T \boldsymbol{l}_j^T \right\|^2, \tag{21}$$

where $\boldsymbol{\sigma}_j$ and $\boldsymbol{l}_j$ are respectively the j^{th} row of $\Delta\boldsymbol{S}_i$ and $\boldsymbol{L}_i$. The detailed derivation is presented in the Appendix.

If the full row rank condition of $\boldsymbol{\Delta R}_i$ is not satisfied, $\text{rank}(\boldsymbol{\Delta R}_i \boldsymbol{\Delta R}_i^T) < n$ and $\boldsymbol{\Delta R}_i \boldsymbol{\Delta R}_i^T$ is singular. Then, instead of (20), we can use Tikhonov regularization:

$$\hat{\boldsymbol{L}}_i = \boldsymbol{\Delta S}_i \boldsymbol{\Delta R}_i^T (\boldsymbol{\Delta R}_i \boldsymbol{\Delta R}_i^T + \lambda \boldsymbol{I})^{-1}. \tag{22}$$

Practically, this implies that one or more inputs motions do not appear in $\boldsymbol{\Delta R}_i$. Therefore, we cannot infer the relationship between these motions and the resulting feature vector changes. In this case it is better to increase M and obtain more data, so that $\boldsymbol{\Delta R}_i$ has full row rank.

Instead of computing the interaction matrix, it is also possible to directly compute its inverse, since this guarantees better control properties, as explained in [26]. With the same data, one can re-write (18) as:

$$\boldsymbol{L}_i^{\oplus} \boldsymbol{\Delta S}_i = \boldsymbol{\Delta R}_i. \tag{23}$$

We can also solve (23) with Tikhonov regularization:

$$\hat{\boldsymbol{L}}_i^{\oplus} = \boldsymbol{\Delta R}_i \boldsymbol{\Delta S}_i^T (\boldsymbol{\Delta S}_i \boldsymbol{\Delta S}_i^T + \lambda \boldsymbol{I})^{-1}. \tag{24}$$

E. Control law and stability analysis

It is now possible to control the robot, using either of the following strategies:

$$\delta\boldsymbol{r}_i = -\alpha \hat{\boldsymbol{L}}_i^{\dagger} (\boldsymbol{s}_i - \boldsymbol{s}_i^*), \tag{25}$$

if one estimates the interaction matrix with (22), where † denotes the pseudo-inverse, or:

$$\delta\boldsymbol{r}_i = -\alpha \hat{\boldsymbol{L}}_i^{\oplus} (\boldsymbol{s}_i - \boldsymbol{s}_i^*) \tag{26}$$

if one estimates the inverse of the interaction matrix with (24). In both equations, $\alpha > 0$ is an arbitrary control gain.

Proposition 1. *Consider that locally, the model (4) closely approximates the interaction matrix $\boldsymbol{L}_i = \hat{\boldsymbol{L}}_i$. For M number of linearly independent displacement vectors $\delta\boldsymbol{r}$ such that the interaction matrix $\hat{\boldsymbol{L}}_i$ is invertible, the update rule (25) asymptotically minimizes the error $\boldsymbol{e}_i = \boldsymbol{s}^* - \boldsymbol{s}_i$.*

Proof: With $\delta\boldsymbol{s}_i = \boldsymbol{s}_{i+1} - \boldsymbol{s}_i$, we can write (4) in discretized form as

$$\boldsymbol{s}_{i+1} = \boldsymbol{s}_i + \boldsymbol{L}_i \delta\boldsymbol{r}_i. \tag{27}$$

The error dynamic

$$\boldsymbol{e}_{i+1} = (1-\alpha)\boldsymbol{e}_i \tag{28}$$

is asymptotically stable for $\alpha \in]0;1[$. This can be proved by considering the Lyapunov function

$$\mathcal{V}(\boldsymbol{e}) = \boldsymbol{e}^T \boldsymbol{e}. \tag{29}$$

Using the error dynamic (28), one can derive:

$$\begin{aligned} \Delta\mathcal{V} &= \mathcal{V}(\boldsymbol{e}_{i+1}) - \mathcal{V}(\boldsymbol{e}_i) \\ &= \boldsymbol{e}_i^T ((1-\alpha)^2 - 1)\boldsymbol{e}_i < 0. \end{aligned} \tag{30}$$

This proves the asymptotic stability of the error $\boldsymbol{e}$ locally by our inputs. ■

F. Model adaptation

Since both the projection matrix $\boldsymbol{U}^T(n)$ and the interaction matrix are local approximations of the full nonlinear mapping, they need to be updated constantly. We choose a receding window approach with the window size equal to M.

At current instance i, we estimate the projection matrix $\boldsymbol{U}_i^T$ and local interaction matrix $\boldsymbol{L}_i$ with M samples of the most recent data. Using the interaction matrix, and the a local target $\boldsymbol{c}_i^*$, we can derive the one-step command $\delta\boldsymbol{r}_i$ by (25). Once we execute the motion $\delta\boldsymbol{r}_i$, a new contour data $\boldsymbol{c}_{i+1}$ is obtained. We move to the next instance $i+1$. A new pair of input and shape data $[\delta\boldsymbol{r}_i, \boldsymbol{c}_{i+1}]$ is obtained. We shift the window by deleting the oldest data in the window and add in the new data pair. Then, using the shifted window, we compute one step control at $i+1$ instance.

The receding window approach ensures that, at each instance, we are using the latest data to estimate the interaction matrix. The overall algorithm is initialized with small random motions around the initial configuration. First, M samples of shape data and the corresponding robot motions are collected. With this initialization, we can simultaneously solve for the projection matrix and estimate the initial interaction matrix using the methods described in Sect. IV-A and IV-D. Using the projection matrix and the initial/desired shapes, we can then find an intermediate target as explained in Sect. IV-C.

We consider quasi-static deformation. Hence, at each iteration the system is in equilibrium and can be linearized according to (4). The data that best captures the current system are the most recent ones. The choice of M is a trade-off between locality and richness. For fast varying deformations[1], we would expect to reduce M as larger M will hinder the locality assumption. Yet, if M is too small, it affects the estimation of $\hat{\boldsymbol{L}}_i$ (refer to detailed discussion in Sect. IV-D).

V. Simulation Results

In this section, we present the numerical simulations that we ran to validate our method.

A. Simulating the objects

We ran simulations on MATLAB (R2018b) with two types of objects: a rigid box and a deformable cable, both constrained to move on a plane. The rigid object is represented by a uniformly sampled rectangular contour. The controllable inputs are its position and orientation. For the cable, we developed a simulator, which is publicly available at https://github.com/Jihong-Zhu/cableModelling2D. The simulator relies on the differential geometry cable model introduced in [27], with the shape defined by solving a constrained optimization problem. The underlying principle is that the object's potential energy is minimal for the object's static shape [28]. Position and orientation constraints – imposed at the cable ends – are input to the simulator. The output is the sampled cable. Figures 4 - 6, 9, 10 show simulated shapes of

[1]The notion of fast or slow varying depends on both the speed of manipulation, and on the objects deformation characteristics (which affect the rate of change in shapes) with regard to the image processing time.

cables and rigid boxes. We choose $K = 50$ samples for both rigid objects and cables. The camera perspective projection is simulated, with optical axis perpendicular to the plane.

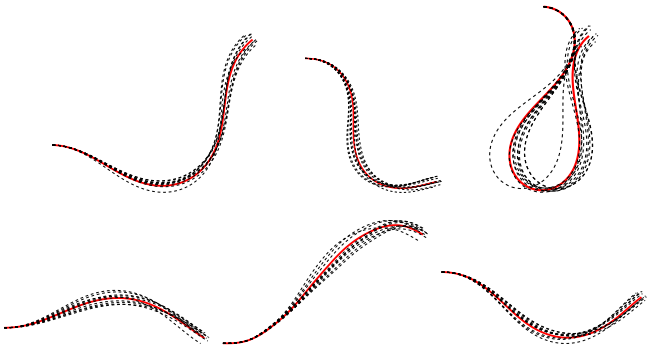


Fig. 4. Six trials conducted to test various choices of feature dimension k for a cable. In each sub-figure, the solid red lines are the initial shapes and the dashed black are the shapes resulting from 10 random motions of the right tip (translations limited to $\pm 5\%$ of the length, rotations limited to $\pm 5^\circ$).

B. Selecting the feature dimension k

To check condition $k = n$ (see Sect. IV-B), we simulate 6 trials with distinct initial shapes of a cable. The dimension of the robot motion vector $\delta \boldsymbol{r}$ is $n = 3$ (two translations and one rotation of the right tip), and the motions are limited (each translation to $\pm 5\%$ of the cable length and the rotation to $\pm 5^\circ$). For each trial, we command $M = 10$ random motions around the initial shape using our simulator. Figure 4 shows the 6 initial cable shapes (solid red) and resulting shapes from 10 random movements (dashed black).

For each trial, we apply PCA to map the cable contour $\boldsymbol{c} \in \mathbb{R}^{2K}$ to feature vector $\boldsymbol{s} \in \mathbb{R}^k$, as explained in Sect. IV-A. We do this for $k = 1$, 2 and 3 and for each of these 18 experiments, we calculate the explained variance $\Upsilon(k)$ with (9). Table I shows these explained variances. In all 6 trials, $k = n = 3$ yields explained variances very close to 1. This result confirms that choosing $k = n$ as the dimension of the feature vector gives an excellent representation of the shape data. It is also possible to select $k = 2$, since the first two components can represent more than 99% of the variance. Nevertheless, the simulation is noise-free. Therefore, although $\Upsilon(k)$ increases little from $k = 2$ to $k = 3$, this increase is not related to noise but to an actual gain in data information.

TABLE I
EXPLAINED VARIANCE $\Upsilon(k)$ FOR THE 6 TRIALS WITH SMALL MOTION.

	trial 1	trial 2	trial 3	trial 4	trial 5	trial 6
$k = 1$	0.727	0.795	0.871	0.847	0.847	0.705
$k = 2$	0.992	0.995	0.996	0.997	0.997	0.994
$k = 3$	0.999	0.999	0.999	0.999	0.999	0.999

At this stage, it is legitimate to ask how does this scale to *larger* movements? Fig. 5 illustrates 10 cable shapes generated by large movements (angle variation: $[-\frac{\pi}{2}, \frac{\pi}{2}]$, maximum translation: 106%). Again, we apply PCA ($M = 10$); Table II shows the $\Upsilon(k)$ resulting from various values of k.

TABLE II
EXPLAINED VARIANCE $\Upsilon(k)$ COMPUTED WITH LARGE MOTION.

k	0	1	2	3	4	5
$\Upsilon(k)$	0	0.5444	0.7218	0.8927	0.9919	0.9990

Comparing Tables I and II, it is noteworthy that $\Upsilon(4)$ with large motion is smaller than $\Upsilon(2)$ with small motion. There are

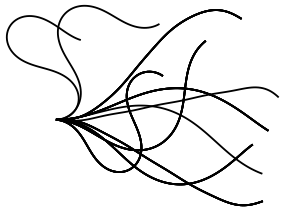


Fig. 5. Ten distinctive cable shapes generated by large motion: angle variation: $[-\frac{\pi}{2}, \frac{\pi}{2}]$, maximum translation: 106% of the cable length.

two possible explanation here. One is that when shapes stays local, the local linear mapping $\boldsymbol{L}$ in (4) remains constant and we need less features to characterize it; the more the shape varies, the more more features we need. Another possible explanation is that for larger motions, $M = 10$ shapes may be insufficient for PCA. Likely, the larger the changes, the larger the number of shapes M needed for proper PCA.

C. Manipulation of deformable objects

With our cable simulator, we can now test the controller to modify the shape from an initial to a desired one. Again, the left tip of the cable is fixed, and we control the right tip with $n = 3$ degrees of freedom (two translations and one rotation). Using the methods described in Sect. IV, we choose window size $M = 5$, the Tikhonov factor $\lambda = 0$, the local target threshold $\eta = 0.8$, the control gain $\alpha = 0.01$. To quantify the effectiveness of our algorithms in driving the contour to $\boldsymbol{c}^*$, we define a scalar measure: the Average Sample Error (ASE). At iteration i, with current contour $\boldsymbol{c}_i$ it is:

$$\text{ASE} = \frac{\|\boldsymbol{c}_i - \boldsymbol{c}^*\|_2}{2K}. \tag{31}$$

A small ASE indicates that the current contour is near the desired one. In Sect. IV-E, we have proved that our controller asymptotically stabilizes the feature vector, $\boldsymbol{s}$ to $\boldsymbol{s}^*$. Hence, since we have also shown that $\boldsymbol{s}$ is a "very good" representation" of $\boldsymbol{c}$, we also expect our controller to drive $\boldsymbol{c}$ to $\boldsymbol{c}^*$, thus ASE to 0. This measure is also used in the real experiments.

Using the cable simulator, we compared the convergence of two control laws proposed in our paper (25) and (26) against a baseline algorithm in [14] which uses Fourier parameters as feature. To make methods compatible, we choose first order Fourier approximation. Note that this results in a feature vector of dimension of 6 (see [14]) which is still twice the number k used in our method. We also normalized the computed control action and then multiplied with the same gain factor 0.01.

Figure 6 shows the evolution of the cable shapes towards the target using (26). Figure 7 compares convergence of our methods against the Fourier-based method. We can observe that our method provides compatible convergence using half the features. Also directly computing of the inverse (26) provides faster convergence than (25).

Taking a step further, we consider the Broyden update law [29], which has been used to update the interaction matrix in classic visual servoing [30]–[32] and shape servoing [33]. Let us hereby show why it is not applicable in our framework.

The Broyden update is an iterative method for estimating

$\boldsymbol{L}_i$ at iteration i. Its standard discrete-time formulation is:

$$\hat{\boldsymbol{L}}_i = \hat{\boldsymbol{L}}_{i-1} + \beta \frac{\delta \boldsymbol{s}_{i-1} - \hat{\boldsymbol{L}}_i \delta \boldsymbol{r}_{i-1}}{\delta \boldsymbol{r}_{i-1}^T \delta \boldsymbol{r}_{i-1}} \delta \boldsymbol{r}_{i-1}^T, \ \forall \boldsymbol{r}_{i-1} \neq \boldsymbol{0} \quad (32)$$

with $\beta \in \,]0;1]$ an adjustable gain. Using our simulator, we estimate the interaction matrix using both Broyden update (with three different values of β) and our receding horizon method (22). We then compare (with $\hat{\boldsymbol{L}}$ estimated with either method) the one-step prediction of the resulting feature vector:

$$\hat{\boldsymbol{s}}_{i+1} = \hat{\boldsymbol{L}}_i \boldsymbol{r}_i + \boldsymbol{s}_i, \quad (33)$$

with the ground truth $\boldsymbol{s}_{i+1}$ from the simulator. The results (plotted in Fig. 8) show that receding horizon outperforms all three Broyden trials. One possible reason is that the components of $\boldsymbol{s}$ fluctuate since (at each iteration) a new matrix $\boldsymbol{U}$ is used. These variations cause the Broyden method to accumulate the result from old interaction matrices, and therefore perform badly on a long term. This result contrasts with that of [33], where the Broyden method performs well since there is a fixed mapping from contour data to feature vector. Another advantage of the receding horizon approach is that it does not require any gain tuning.

D. Manipulation of rigid objects

The same framework can also be applied to rigid object manipulation. Consider the problem of moving a rigid object to a certain position and orientation via visual feedback. This time, the shape of the object does not change, but its pose will (it can translate and rotate). We use the same M, λ, η and α as for cable manipulation. We compare the convergence of two control laws proposed in our paper (25) and (26) against a baseline using image moments [17]. The translation and orientation can be represented with image moments and the analytic interaction matrix can be computed as explained in [17]). To make methods compatible, we normalize the computed control and then multiply it by the same factor 0.01.

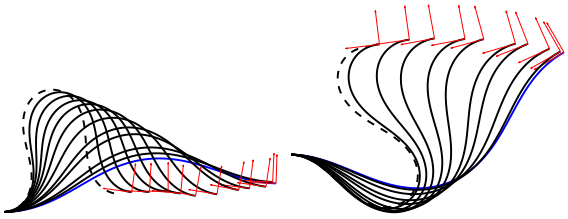


Fig. 6. Cable manipulation with a single end-effector, moving the right tip. The blue and black lines are the initial and intermediate shapes, respectively, and the dashed black line is the target shape. The red frame indicates the end-effector position and orientation generated by our controller.

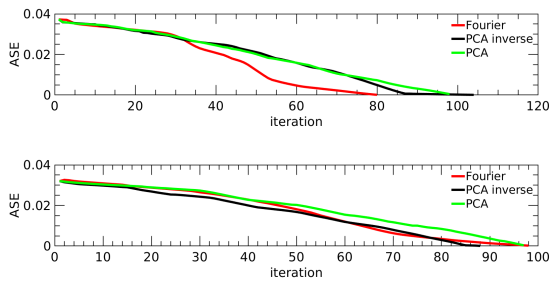


Fig. 7. The evolution of ASE of the simulated cable manipulation using our method against the Fourier-based method as baseline. Top: left simulation in Fig. 6. Bottom: right simulation in Fig. 6.

Figure 9 shows two simulations where our controller successfully moves a rigid object from an initial (blue) to a desired (dashed black) pose using control law (26). Figure 11 compares convergence of our methods against the image moments method. We can observe that our method provides a slightly slower convergence. Directly computing the inverse (26) provides a convergence similar to (25). Later, we will show why our method is slower. Yet, the fact that it can be applied on both deformable and rigid objects makes it stand out over the other techniques.

Taking a step further, we would like to analyze locally what does each component of the feature vector represent. To this end, we apply $M = 10$ random movements (rotation range $[-0.11, 0.09]$, maximum translation 15% of the width) to a rigid rectangular object of length 5 times larger than width (see Fig. 10). We compute the projection matrix as explained in Sect. IV-A, and transform the contour samples to feature vectors. Following the rationale explained in Sect. IV-B, we set $k = n = 3$. Then, we seek the relationship – at each iteration – between the object pose x, y, θ and the components of the feature vector $\boldsymbol{s}$ generated by PCA. To this end, we use bivariate correlation [34] defined by:

$$\rho = \frac{E[(\xi - \bar{\xi})(\zeta - \bar{\zeta})]}{\sigma_\xi \sigma_\zeta}, \quad (34)$$

where ξ and ζ are two variables with expected values $\bar{\xi}$ and $\bar{\zeta}$ and standard deviations σ_ξ and σ_ζ. An absolute correlation $|\rho|$ close to 1 indicates that the variables are highly correlated. All the simulations in Fig. 10 exhibit similar correlation between the computed feature vector and the object pose. In Table III, we show one instance (top left simulation in Fig. 10) of the correlation between variables, with high absolute correlations marked in red. It is clear from the table that each component in the feature vector relates strongly to one pose parameter. We further demonstrate the correlation in Fig. 12, where we plot the evolution of object poses and feature components. Note that s_2 and θ are negatively correlated. The slower convergence (compared to image moments) could be as a result of non-unitary correlation between extracted features and object pose.

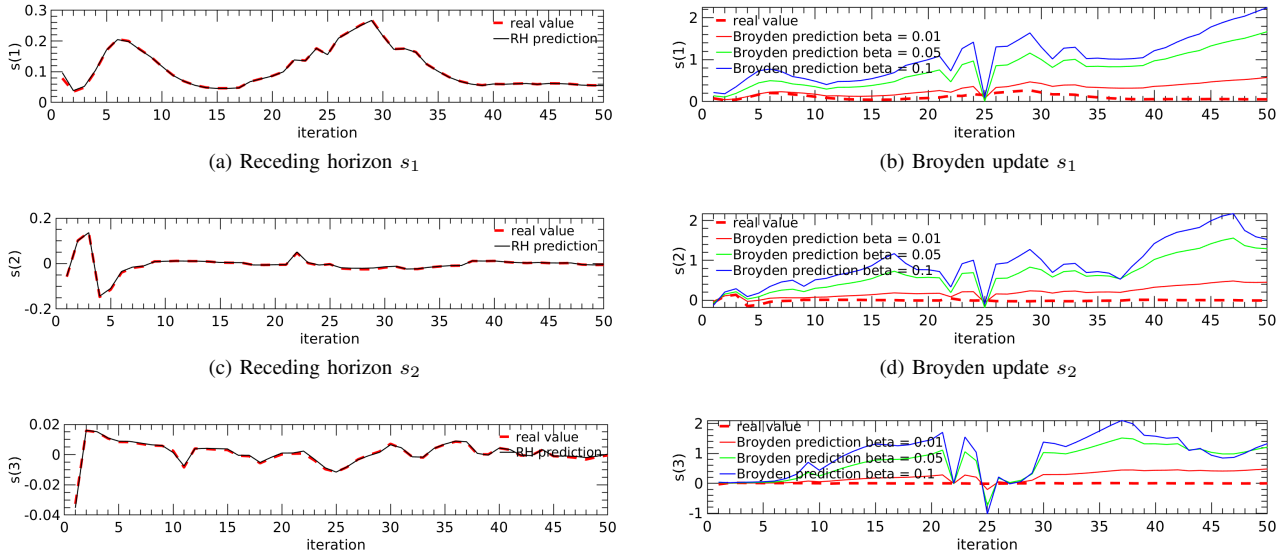
TABLE III
CORRELATION ρ BETWEEN s_1, s_2, s_3 AND x, y, θ.

	x	y	θ
s_1	-0.2819	-0.3343	0.9887
s_2	0.2607	-0.8547	-0.0465
s_3	0.9230	0.3629	-0.1426

VI. EXPERIMENTS

Figure 13 outlines our experimental setup. We use a KUKA LWR IV arm. We constrain it to planar ($n = 3$) motions $\delta \boldsymbol{r}$, defined in its base frame (red in the figure): two translations δx and δy and one counterclockwise rotation $\delta\theta$ around z. A Microsoft Kinect V2 observes the object[2]. A Linux-based 64-bit PC processes the image at 30 fps. In the following sections, we first introduce the image processing for contour extraction, then present the experiments.

[2]We only use the RGB image – not the depth – from the sensor.



(a) Receding horizon s_1

(b) Broyden update s_1

(c) Receding horizon s_2

(d) Broyden update s_2

(e) Receding horizon s_3

(f) Broyden update s_3

Fig. 8. Comparison – for estimating $\boldsymbol{s}$ – of the receding horizon approach (RH, left) and of the Broyden update (right, with three values of β). The topmost, middle and bottom plots show the one step prediction of s_1, s_2 and s_3, respectively. In all plots, the dashed red curve is the ground truth from the simulator. The plots clearly show that the receding horizon approach outperforms all three Broyden trials.

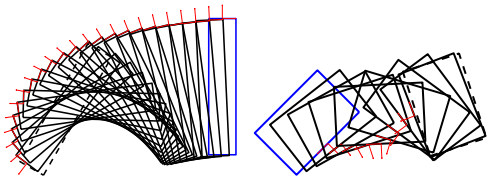


Fig. 9. Manipulation of a rigid object with a single end-effector (red frame). The initial, intermediate and desired contours are respectively blue, solid black and dashed black. Note that in both cases, our controller moves the object to the desired pose.

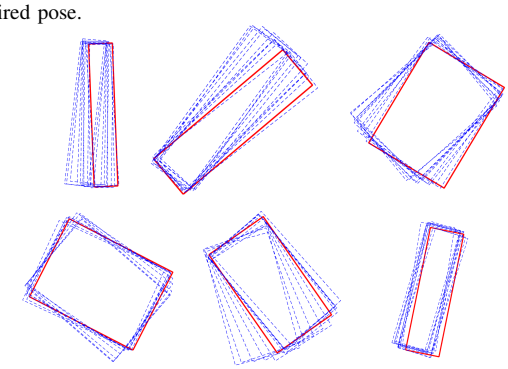


Fig. 10. From an initial (red) pose, we generate 10 (dashed blue) random motions of a rigid object.

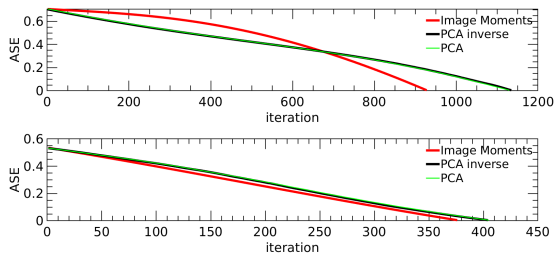


Fig. 11. Evolution of ASE of the simulated rigid object manipulation using our method against image moments. Top: left simulation in Fig. 9, Bottom: right simulation in Fig. 9.

A. *Image processing*

This section explains how we extract and sample the object contours from an image. We have developed two pipelines, according to the kind of contours (See Fig. 14): open (e.g., representing a cable) and closed. We hereby describe the two.

1) Open Contours: The overall pipeline for extracting an open contour is illustrated in Fig. 15 and **Algorithm 2**. On the initial image, the user manually selects a Region of Interest (ROI, see Fig. 15a) containing the object. We then apply thresholding, followed by a morphological opening, to the

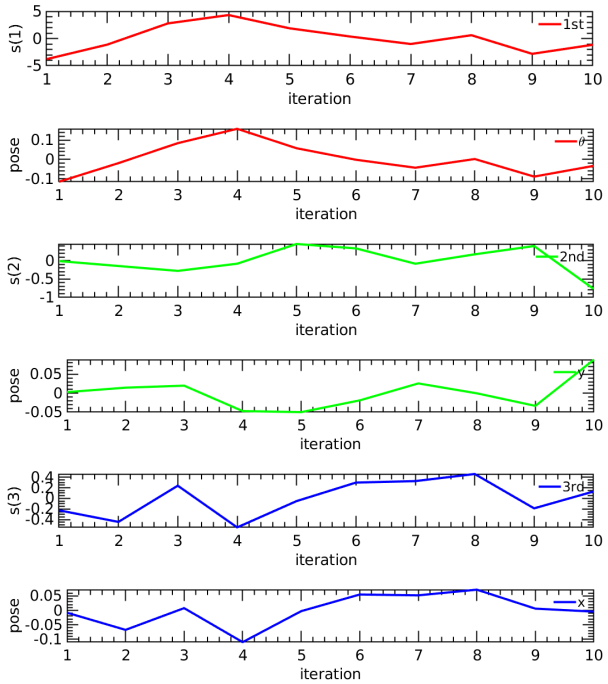


Fig. 12. Progression of the auto-generated feature components (row 1, 3, 5: s_1, s_2, s_3) vs. object pose (row 2, 4, 6: x, y, θ). We have purposely arranged the variables with high correlation with the same color.

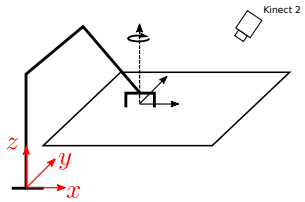


Fig. 13. Overview of the experimental setup.

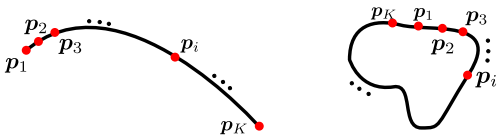
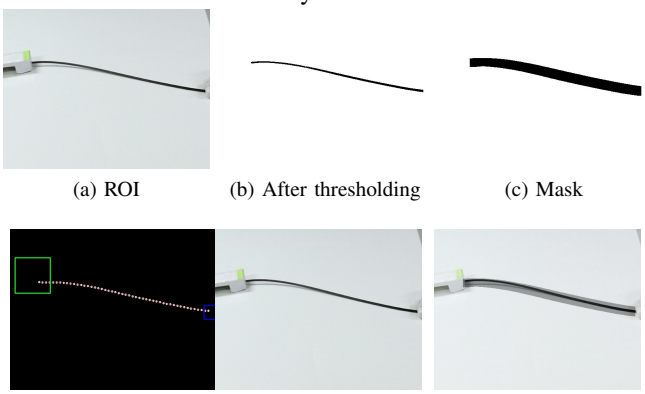


Fig. 14. Open (left) and closed (right) contours can be both represented by a sequence of sample pixels in the image.

image, to remove the noise and obtain a binary image as in Fig. 15b. This image is dilated to generate a mask (Fig. 15c), which is used as the new ROI to detect the object on the following image. Figure 15e is the object after a small manipulation motion and 15f shows the mask with the grey color which contains the cable in Fig. 15e. On each binary image, we apply **Algorithm 2** to uniformly sample the object (see Fig. 15d, where the green box indicates the end-effector). This is the contour $\boldsymbol{c}$ used by our controller.

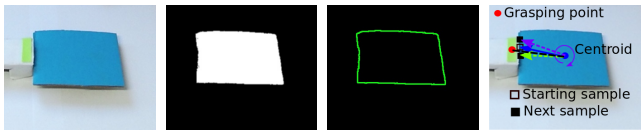


(a) ROI (b) After thresholding (c) Mask

(d) Sampled data (e) Next image (f) Cable in the mask

Fig. 15. Image processing steps needed to obtained the sampled open contour of an object (here, a cable).

2) Closed Contours: The procedure is shown in Fig. 16. For an object with uniform color (in the experiment blue), we apply HSV segmentation, followed by Gaussian blur of size 3, and finally the OpenCV *findContour* function, to get the object contour. The contour is then re-sampled using **Algorithm 2**. The starting point and the order of the samples is determined by tracking the grasping point (red dot in Fig. 16d) and the centroid of the object (blue dot). We obtain the vector connecting the grasping point to the centroid. Then, the starting sample is the one closest to this vector, and we proceed along the contour clockwise.



(a) (b) (c) (d)

Fig. 16. Image processing for getting a sampled closed contour: (a) original image, (b) image after thresholding and Gaussian blur, (c) extracted contour, (d) finding the starting sample and the order of the samples.

Algorithm 2 Generate fixed number of sampled data
Input: $\boldsymbol{P}' = [\boldsymbol{p}'_1, \cdots, \boldsymbol{p}'_L]$, original ordered sampled data and K, desired number of data samples generated
Output: $\boldsymbol{P} = [\boldsymbol{p}_1, \cdots, \boldsymbol{p}_K]$, re-sampled data.

compute the full length of $\mathcal{L}$ of $\boldsymbol{P}'$.
compute desired distance per sample: $\mu = \mathcal{L}/K$
$k = 1,\ h = 1$
$\text{dist} = 0$
$\boldsymbol{p}_k = \boldsymbol{p}'_l$
while $k \leq K$ **do**
  $\boldsymbol{p}_{\text{next}} = \boldsymbol{p}'_{h+1}$
  $d = ||\boldsymbol{p}'_{k+1} - \boldsymbol{p}_{\text{next}}||_2$
  if $d + \text{dist} \leq \mu$ **then**
    $\text{dist} = \text{dist} + d$
    $\boldsymbol{p}_{k+1} = \boldsymbol{p}_{\text{next}}$
    $k = k + 1,\ h = h + 1$
  else
    $\boldsymbol{p}_{k+1} = \boldsymbol{p}_k + \mu \frac{\boldsymbol{p}'_{k+1} - \boldsymbol{p}_{\text{next}}}{d}$
    $k = k + 1$
    $\text{dist} = 0$
  end if
end while

B. Vision-based manipulation

In this section, we present the experiments that we ran to validate our algorithms, also visible at https://youtu.be/gYfO2ZxZ5KQ. To demonstrate the generality of our framework, we tested it with:

- Rigid objects represented by closed contours,
- Deformable objects represented by open contours (cables),
- Deformable objects represented by closed contours (sponges).

We carried out different experiments with a variety of initial and desired contours and camera-to-object relative poses. The variety of both geometric and physical properties demonstrates the robustness of our framework. The variety of camera-to-object relative poses shows that—as usual in image-based visual servoing [1]—camera calibration is unnecessary. The algorithm and parameters are the same in all experiments; the only differences are in the image processing, depending on the type of contour (closed or open, see Sect. VI-A).

We obtain the desired contours by commanding the robot with predefined motions. Once the desired contour is acquired, the robot goes back to the initial position, and then should autonomously reproduce the desired contour. Again, we set the number of features $k = n = 3$, and use $K = 50$ samples to represent the contour $\boldsymbol{c}$. We set the window size $M = 5$, both for obtaining the feature vector $\boldsymbol{s}$ and the interaction matrix $\boldsymbol{L}$. The control gain is 0.01, the local target threshold $\eta = 0.8$ and the Tikhonov factor $\lambda = 0.01$. At the beginning of each experiment, the robot executes 5 steps of small[3] random motions to obtain the initial features and interaction matrix.

[3]The notion of small is relative, and usually dependent on the size of the object the robot is manipulating. Refer to Sect. IV-A (especially Fig. 4) for a discussion on this.

For all the experiments, we set the same termination condition at iteration $i+1$ using ASE defined in (31) such that:

1) $\text{ASE}_i < 1$ pixel and
2) $\text{ASE}_{i+1} \geq \text{ASE}_i$.

In the graphs that follow, we show the evolution of ASE in blue before the termination condition, and in red after the condition (until manual stop by the operator).

Figure 17 presents 8 experiments, one per column. Columns 1 – 3, 4 – 6 and 7 – 8 show respectively manipulation of: cable, rigid object and sponge. The first row presents the full RGB image obtained from Kinect V2. The second and third rows zoom in on the manipulation at the initial and final iterations. We track the end-effector in the image with a green marker for contour sampling. The desired and current contours are drawn in red and blue, respectively.

Figure 18 shows the decreasing trend of error ASE for each experiment. The initial increase of ASE in the experiments can be due to the random motion at the beginning of the experiments. In general, we found that ASE is more noisy for the closed than for the open contour. This discontinuity is visible in Figures 18c and 18d (zigzag evolution). Such noise is likely introduced by the way we sampled the contour. When we have false contour data, the value of ASE may encounter a sudden discontinuity. Figure 19 shows examples of these false samples, output by the image processing pipeline. Despite these errors, thanks to the "forgetting nature" of the receding horizon and to the relatively small window size ($M = 5$), the corrupted data will soon be forgotten, and it will not hinder the overall manipulation task. Yet, the overall framework would benefit from a more robust sensing strategy, as in [35].

Finally, since our framework can deal with both rigid and deformable objects, we tested it in two experiments where the same object (a sponge) can be both rigid (in the free space), and deformed (when in contact with the environment). These experiments require the robot to: 1) move the object, establish contact, 2) give the object the desired shape, by relying on the contact. Figure 20 presents these two original "move and shape" servoing experiments with the corresponding errors ASE plotted in Fig. 21. We use a second fixed robot arm to generate the deforming contact. As the figures and curves show, both experiments were successful.

The success of the "move and shape" task is largely dependent on the contact establishment. However, even when the initial contact has some misalignment (see Fig. 20c and 20g), our framework can still reduce the ASE to give a reasonable final configuration (see Fig. 20h and Fig. 21b).

VII. Conclusion

In this paper, we propose algorithms to automatically and concurrently generate object representations (feature vectors) and models of interaction (interaction matrices) from the same data. We use these algorithms to generate the control inputs enabling a robot to move and shape the said object, be it rigid or deformable. The scheme is validated with comprehensive experiments, including a desired contour that requires both moving and shaping. We believe it is unprecedented in previous research.

Our framework adopts a model-free approach. The system characteristics are computed online with visual and manipulation data. We do not require camera calibration, nor a priori knowledge of the camera pose, object size or shape. An open question remains the management of 6 DOFs motions of arms. Indeed, while the proposed control strategy can be easily generalized to 6 DOFs motions, it relies on a sufficiently accurate extraction of feature vectors from vision sensors. A very challenging task is to generate complete 3D feature vectors of objects from a limited sensor set, due to partial view of objects and occlusions.

The framework could benefit from robust sensing of deformation. In fact, one obvious setback is that the representation and model of interaction are extremely local. Thus, they cannot guarantee global convergence. In addition, our framework cannot infer whether a shape is reachable or not. This drawback is solvable by using a global deformation model for control. But as we mentioned earlier, a global model usually requires an offline identification phase which we want to avoid. In fact, for different objects, we will need to re-identify the model. There is a dilemma in using a global deformation model.

Maybe one of the possible solutions to this dilemma is to have both our method and deep learning based methods running in parallel. While our scheme enables fast online computation and direct manipulation, the extracted data can be used by a deep neural network to obtain a global interaction mapping. Once a global mapping is learned, it can later be used for direct manipulation and to infer feasibility of the goal shape.

References

[1] 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.

[2] H. Inoue, "Hand-eye coordination in rope handling," in *Proc. of Int. Symposium on Robotics Research*. MIT PRESS, 1984, pp. 163–174.

[3] P. W. Smith, N. Nandhakumar, and A. K. Ramadorai, "Vision based manipulation of non-rigid objects," in *IEEE Int. Conf. on Robotics and Automation*, vol. 4. IEEE, 1996, pp. 3191–3196.

[4] D. Berenson, "Manipulation of deformable objects without modeling and simulating deformation," in *2013 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, 2013, pp. 4525–4532.

[5] R. Lagneau, A. Krupa, and M. Marchal, "Active deformation through visual servoing of soft objects," in *IEEE Int. Conf. on Robotics and Automation*, 2020.

[6] D. Navarro-Alarcon, Y. Liu, J. G. Romero, and P. Li, "On the visual deformation servoing of compliant objects: Uncalibrated control methods and experiments," *Int. Journal of Robotics Research*, vol. 33(11), pp. 1462–1480, 2014.

[7] D. Navarro-Alarcon and Y.-H. Liu, "Fourier-based shape servoing: A new feedback method to actively deform soft objects into desired 2d image contours," *IEEE Trans. on Robotics*, vol. 34(1), pp. 272–279, 2018.

[8] D. Navarro-Alarcon, A. Cherubini, and X. Li, "On model adaptation for sensorimotor control of robots," in *Chinese Control Conference*, 2019.

[9] A. Cherubini, V. Ortenzi, A. Cosgun, R. Lee, and P. Corke, "Model-free vision-based shaping of deformable plastic materials," *The Int. Journal of Robotics Research (to appear)*, 2020.

[10] Z. Wang, X. Li, D. Navarro-Alarcon, and Y.-h. Liu, "A unified controller for region-reaching and deforming of soft objects," in *2018 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*.

[11] D. Navarro-Alarcon and Y. Liu, "Uncalibrated vision-based deformation control of compliant objects with online estimation of the jacobian matrix," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2013.

[12] M. Laranjeira, C. Dune, and V. Hugel, "Catenary-based visual servoing for tethered robots," in *2017 IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 732–738.

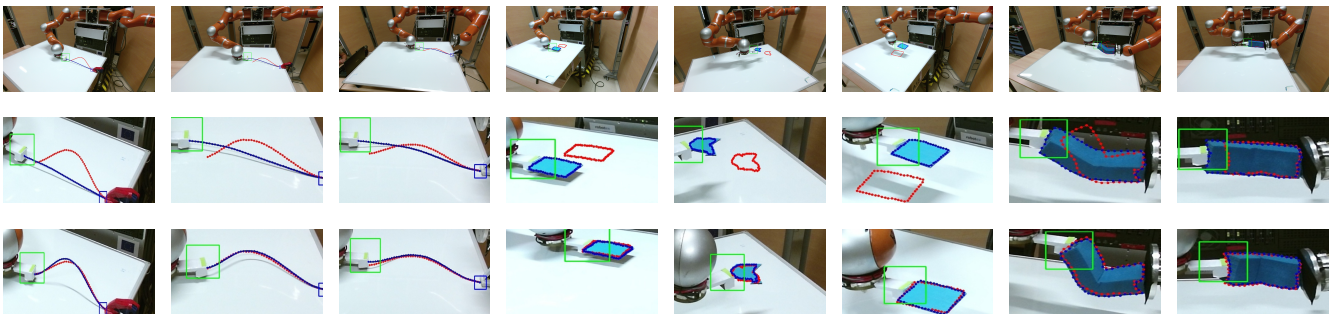
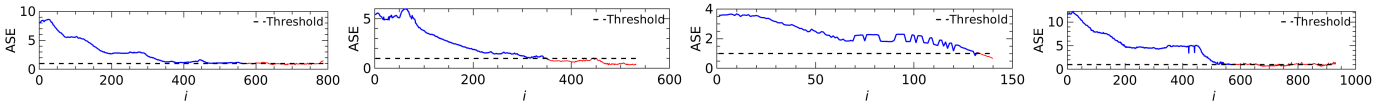


Fig. 17. Eight experiments with the robot manipulating different objects. From left to right: a cable (columns 1 – 3), a rigid object (columns 4 – 6) and a sponge (columns 7 and 8). The first row shows the full Kinect V2 view, and the second and the third columns zoom in to show the manipulation process at the first and last iterations. The red contour is the desired one, whereas the blue contour is the current one. The green square indicates the end-effector.



(a) cable – column 1 (b) cable – column 2 (c) cable – column 3 (d) rigid object – column 4

(e) rigid object – column 5 (f) sponge – column 6 (g) sponge – column 7 (h) sponge – column 8

Fig. 18. Evolution of e_i at each iteration i, for the 8 experiments of Fig. 17. The black dashed lines indicate the threshold ASE $= 1$ pixel. The blue curves show e_i until the termination condition, whereas the red curves show the error until manual termination by the human operator.

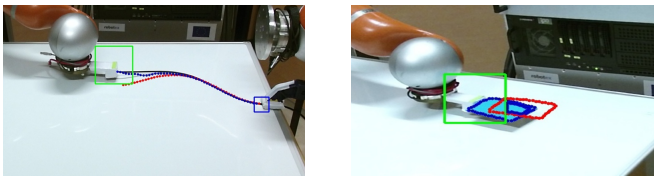
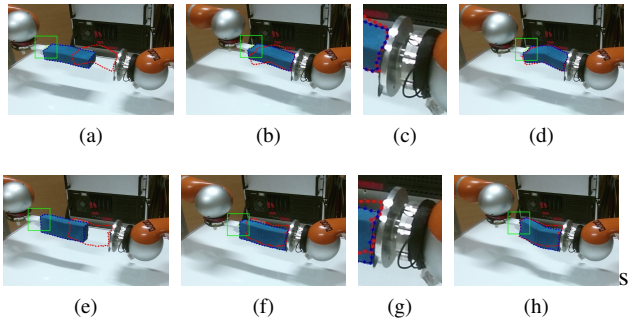


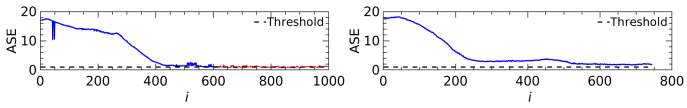
Fig. 19. False contour data from the image can cause noise in ASE.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 20. Two "move and shape" experiments grouped into two rows. The desired contour (red dotted) is far from the initial one. This requires the robot to 1) move the object, establish contact with the right – fixed – robot arm, 2) give the object the desired shape, by relying on the contact. The first column shows the starting configuration, the second column presents the contact establishment, and the third column zooms in to show the alignment. The last column shows the final results.



(a) move and shape – row 1 (b) move and shape – row 2

Fig. 21. The evolution of e_i for the experiments of Fig. 20. The black dashed line indicates the threshold ASE $= 1$. The blue curves show e_i until the termination condition, whereas the red curves show the error until manual termination by the human operator.

[13] ——, "Catenary-based visual servoing for tether shape control between underwater vehicles," *Ocean Engineering*, vol. 200, pp. 1–19, 2020.

[14] J. Zhu, B. Navarro, P. Fraisse, A. Crosnier, and A. Cherubini, "Dual-arm robotic manipulation of flexible cables," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2018.

[15] A. Nair, D. Chen, P. Agrawal, P. Isola, P. Abbeel, J. Malik, and S. Levine, "Combining self-supervised learning and imitation for vision-based rope manipulation," in *IEEE Int. Conf. on Robotics and Automation*, 2017.

[16] Z. Hu, T. Han, P. Sun, J. Pan, and D. Manocha, "3-d deformable object manipulation using deep neural networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4255–4261, 2019.

[17] F. Chaumette, "Image moments: a general and useful set of features for visual servoing," *IEEE Trans. on Robotics*, vol. 20(4), pp. 713–723, 2004.

[18] C. Collewet, E. Marchand, and F. Chaumette, "Visual servoing set free from image processing," in *2008 IEEE Int. Conf. on Robotics and Automation*. IEEE, 2008, pp. 81–86.

[19] C. Collewet and E. Marchand, "Photometric visual servoing," *IEEE Trans. on Robotics*, vol. 27, no. 4, pp. 828–834, 2011.

[20] M. Bakthavatchalam, F. Chaumette, and E. Marchand, "Photometric moments: New promising candidates for visual servoing," in *2013 IEEE Int. Conf. on Robotics and Automation*. IEEE, 2013, pp. 5241–5246.

[21] E. Marchand, "Subspace-based direct visual servoing," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2699–2706, 2019.

[22] S. K. Nayar, S. A. Nene, and H. Murase, "Subspace methods for robot vision," *IEEE Trans. on Robotics and Automation*, vol. 12(5), pp. 750–758, 1996.

[23] K. Deguchi and T. Noguchi, "Visual servoing using eigenspace method and dynamic calculation of interaction matrices," in *Proc. of 13th IEEE Int. Conf. on Pattern Recognition*, 1996.

[24] Q. Sang and G. Tao, "Adaptive control of piecewise linear systems: the state tracking case," *IEEE Trans. Autom. Control*, vol. 57, no. 2, pp. 522–528, Feb 2012.

[25] G. Strang, *Introduction to linear algebra*. Wellesley-Cambridge Press Wellesley, MA, 1993, vol. 3.

[26] J.-T. Lapresté, F. Jurie, M. Dhome, and F. Chaumette, "An efficient method to compute the inverse jacobian matrix in visual servoing," in *IEEE Int. Conf. on Robotics and Automation*, 2004.
[27] H. Wakamatsu and S. Hirai, "Static modeling of linear object deformation based on differential geometry," *The Int. Journal of Robotics Research*, vol. 23, no. 3, pp. 293–311, 2004.
[28] H. Wakamatsu, S. Hirai, and K. Iwata, "Modeling of linear objects considering bend, twist, and extensional deformations," in *IEEE Int. Conf. on Robotics and Automation*, 1995.
[29] C. G. Broyden, "A class of methods for solving nonlinear simultaneous equations," *Mathematics of computation*, vol. 19(92), pp. 577–593, 1965.
[30] K. Hosoda and M. Asada, "Versatile visual servoing without knowledge of true jacobian," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 1994.
[31] M. Jagersand, O. Fuentes, and R. Nelson, "Experimental evaluation of uncalibrated visual servoing for precision manipulation," in *Int. Conf. on Robotics and Automation*, 1997.
[32] F. Chaumette and S. Hutchinson, "Visual servo control, part II: Advanced approaches," *IEEE Robotics and Automation Magazine*, vol. 14, no. 1, pp. 109–118, 2007.
[33] D. Navarro-Alarcon, Y. Liu, J. G. Romero, and P. Li, "Model-free visually servoed deformation control of elastic objects by robot manipulators," *IEEE Trans. on Robotics*, vol. 29(6), pp. 1457–1468, 2013.
[34] W. Feller, *An introduction to probability theory and its applications*. John Wiley & Sons, 2008, vol. 2.
[35] C. Chi and D. Berenson, "Occlusion-robust deformable object tracking without physics simulation," in *2019 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*.

APPENDIX

We hereby provide the proof of equivalence between (20) and the solution to the optimization problem mentioned in Sect. IV-D. Without loss of generality, we omit index i.

Given $\boldsymbol{\Delta S} \in \mathbb{R}^{n\times M}$ and $\boldsymbol{\Delta R} \in \mathbb{R}^{n\times M}$, we would like to find $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ such that:

$$\boldsymbol{\Delta S} = \boldsymbol{L}\boldsymbol{\Delta R}. \tag{35}$$

We can solve (35) by the optimization problem $\min_{\boldsymbol{L}}(\mathcal{J})$ with

$$\mathcal{J} = \sum_{j=1}^{n} ||\delta\boldsymbol{\sigma}_j^T - \Delta\boldsymbol{R}^T\boldsymbol{l}_j^T||^2, \tag{36}$$

where $\boldsymbol{\sigma}_j$ and $\boldsymbol{l}_j$ are respectively the j^{th} row of $\Delta\boldsymbol{S}$ and $\boldsymbol{L} \in \mathbb{R}^{n\times n}$, and $||\cdot||_2$ denotes the Euclidean 2-norm. We want to show that the solution of the optimization problem is equivalent to (20), that is:

$$\hat{\boldsymbol{L}} = \boldsymbol{\Delta S}\boldsymbol{\Delta R}^T(\boldsymbol{\Delta R}\boldsymbol{\Delta R}^T)^{-1}. \tag{37}$$

As in Section IV-D, we assume that $\text{rank}(\boldsymbol{\Delta R}) = n$. We can decompose the solving of $\min_{\boldsymbol{L}}(\mathcal{J})$ into solving individual sub-problems. We rewrite (36) as:

$$\min_{\boldsymbol{L}}(\mathcal{J}) = \min_{\boldsymbol{l}_1}(\mathcal{J}_1) + \min_{\boldsymbol{l}_2}(\mathcal{J}_2) + \ldots + \min_{\boldsymbol{l}_n}(\mathcal{J}_n), \tag{38}$$

where each $\mathcal{J}_i$ is

$$\mathcal{J}_i = ||\delta\boldsymbol{\sigma}_j^T - \Delta\boldsymbol{R}^T\boldsymbol{l}_j^T||_2. \tag{39}$$

The individual optimization problem

$$\min_{\boldsymbol{l}_i}(\mathcal{J}_i) = \min_{\boldsymbol{l}_i} ||\delta\boldsymbol{\sigma}_j^T - \Delta\boldsymbol{R}^T\boldsymbol{l}_j^T||_2 \tag{40}$$

is a standard linear least square problem:

$$\Delta\boldsymbol{R}^T\boldsymbol{l}_j^T = \delta\boldsymbol{\sigma}_j^T. \tag{41}$$

Its solution is:

$$\boldsymbol{l}_j^T = (\boldsymbol{\Delta R}\boldsymbol{\Delta R}^T)^{-1}\boldsymbol{\Delta R}\delta\boldsymbol{\sigma}_j^T. \tag{42}$$

Transposing (42) on both sides, we have:

$$\boldsymbol{l}_j = \delta\boldsymbol{\sigma}_j\boldsymbol{\Delta R}^T(\boldsymbol{\Delta R}\boldsymbol{\Delta R}^T)^{-1}, \tag{43}$$

and writing (43) in matrix form:

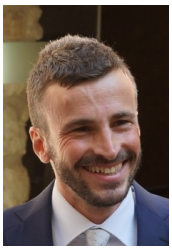
$$\hat{\boldsymbol{L}} = \boldsymbol{\Delta S}\boldsymbol{\Delta R}^T(\boldsymbol{\Delta R}\boldsymbol{\Delta R}^T)^{-1}, \tag{44}$$

which corresponds to (37) ■.

Jihong Zhu received the M.Sc. in Systems and Control in 2015 from TU Delft and the Ph.D. in Robotics from University of Montpellier in 2020. He conducted his doctoral research at LIRMM. In 2019, he visited ROMI Lab hosted by Dr. David Navarro-Alarcon for 4 months. He is currently a postdoc at Cognitive Robotics department, TU Delft. His research interests include: sensor-based control, and manipulation of soft objects.

David Navarro-Alarcon (GS'06–M'14–SM'19) received the Ph.D. degree in mechanical and automation engineering from The Chinese University of Hong Kong, NT, Hong Kong, in 2014. He was Research Assistant Professor at the CUHK Robotics Institute, from 2015 to 2017. Since 2017, he has been with The Hong Kong Polytechnic University, where he is Assistant Professor at the Department of Mechanical Engineering. His current research interests include perceptual robotics and control theory.

Robin Passama received the M.Sc. in Computer Science in 2002 and graduated in 2006 with a PhD Degree in Computer Science, from the University of Montpellier. As a research engineer, he then worked on many projects and contributed in various experiments. He became a permanent CNRS member in 2012. His main area of expertise is software engineering for robotics.



Andrea Cherubini received the M.Sc. in Mechanical Engineering in 2001 from the University of Rome La Sapienza and a second M.Sc. in Control Systems in 2003 from the University of Sheffield, U.K. In 2008, he received the Ph.D. in Control Systems from La Sapienza. From 2008 to 2011, he was postdoc at INRIA Rennes. Since 2011, he is Associate Professor at Université de Montpellier. His research interests include: physical humanrobot interaction, and manipulation of soft objects.