

Response to reviewers' comments

Reviewer 1

In this revision, the authors solved the initial problems on article structure, citations, figures, and contents.

We thank the reviewer for their confirmation that we have satisfied all their comments, and would like to thank them again for their remarks, which significantly improved our paper.

Reviewer 2

Page 1. Last Paragraph. There is an enumeration of novel contributions, you mentioned the First, and the Third, to avoid any confusion I would recommend to remark which is the second.

We apologise, this was a typo. We have listed two main technical contributions.

Page 1-2. It is possible to infer the meaning of FEM, but in order to avoid any possible confusion you might indicate its meaning explicitly when the concept is first used.

We have fixed this.

I find Fig 2, 3, and 4 difficult to read, however I don't consider they should be necessarily changed.

We agree this was difficult to interpret. To aid the reader we have added some text descriptions for the most important elements to make the diagram more readable.

At the end of subsection VII. B, when the authors refer to "all elastic and plastic terms" might be illustrative to add references to the corresponding equations.

Vero?

Within the Model Overview, it would be interesting if the authors would briefly highlight the advantage of having generative models for those who are not vastly familiarized with the topic.

Thank you for this suggestion. We have added a short explanation within the section Model Overview. This explains that generative models are appealing because they allow us to solve both prediction and classification problems, whereas discriminative models are typically only suitable for classification.

When interpreting results what are the consequences when triangles in the border of the mesh overlap each other? Is there any consequence that might be significant to the result interpretation?

Vero?

In section V, during the training the segmentation assumes that the color of the object is known. Therefore, when you say that the algorithm could learn and classify previously unknown materials, at least the machine must partially know the material. Also the authors mention that the simplified active contours approach is not the most accurate, What are the consequences of this approximation? Regarding to the chosen integration time step, does the results vary considerably when 0.01 is considered?

Vero?

In section VII.E, Table III, Is there any special reason to have the same result for the training and test1 sets in the case of the sponge? It would be better if the authors offer deeper insights regarding the interpretation of the figures shown in Table III.

Vero?

Reviewer 3

Reading the abstract and the introduction, I was expecting a framework in which the robot creates a representation of the environment (in this case the two objects) while it is interacting with it (looking at and pushing the object).

We are sorry that this set false expectations. We have altered the wording in several places to make the reader aware that we are not tackling this problem on-line, but off-line. We have also read and noted the references you mention in the introduction.

While the authors recognize the issue of the on-line learning in the introduction (page 1, col 2, line 35), they do not address that in the paper.

From your comments, we believe that mentioning on-line learning, while important, was misleading in terms of our paper, so we have withdrawn mention of it.

Also they refer the problem of getting access to partial views of the objects but it seems that in this paper the object is viewed from a convenient point of view.

We have added text to make clear the restriction in our study of using a single view chosen by the user.

Finally, the word “learning from data” suggests to me that the model acquires a representation of the environment based on some kind of non-parametric model, such as neural network. However, the models seem carefully designed to represent the object deformation and the stress-gain diagram. Even if the difference between learning and calibration is subtle, in my opinion, the use of the word “learning” creates an expectation that is not fulfilled in the rest of the paper.

We have changed the wording in the abstract to note that we calibrate the models from data. We hope this makes it clearer where on the spectrum of parameter tuning to full autonomous learning we lie. We now emphasise that we perform off-line learning, we sometimes refer to it as calibration, and we mention that we don't address the issues of on-line learning. We hope this changes the tone of the introduction enough so as not to mislead.

What happen if we add a new material? Do we need to add a new feature space?

If the object is plastic or elastic then we don't need to add a new feature space, but we would be wise to calibrate a new model. We have added a comment to this effect in the conclusion.

I would add a figure like Fig. 9a at this point. In this way the reader can have an idea of the setup, how vision acquire information and how the mass-spring model depends on the visual information itself. For example, it might be interesting to combine Fig 9a and Fig 2 in single sketch of the model.

Thank you for this suggestion. We have prepared a new version of Figure 2, as you suggest. This incorporates elements from the old Figure 2 and Figure 9.

A consideration about the organisation of the following sections. Given that the framework consists of two models, force prediction and shape prediction, I would expect section III.1 to be Force prediction and III.2 to the shape prediction (or vice-versa). However, the organization of Sections IV, V, VI, in my opinion is misleading

We have reorganised the paper as you request.

Finally, I would start describing the algorithm from the visual system. It seems that the visual system has a very important role in this paper but it is not highlighted. Indeed, the mass-spring model described in section IV should be created within the perimeter of the segmented object. Thus, it requires a visual system capable of localizing and segmenting the object. In my understanding, this information is necessary to initialize the position of the particles (described in section IV) inside the object. Regarding this point, I

would like to know how the particles are placed inside the object. What happens if we change the size of the object? Do we need to change the number of particles? What happens if the object is not rectangular?

Vero?

Eq. 19. I really do not understand how the force could be proportional to the velocity. This point should be clarified. My understanding is that the authors assumed the acceleration (a) to be constant during the interval h , so that $h \cdot F = h \cdot a \cdot m \Rightarrow F = v \cdot m / h$. Given that h is a constant, the force (F), the velocity (v) or the mass (m) are recovered with a scale factor. Is that correct? If so, what is the difference between this approach and the quasi-static approach proposed by Frank. Please, explain.

Vero?

This section is quite long and I think it could be simplified notably. The authors could simply state the energy $E = \text{Sum}(i=1:N) [0.5 k_i \cdot C_i(p1, pn)^2]$, where N is the number of constraints. There are several constraints to preserve the length, the area, etc. and the force is given as negative gradient of the potential energy with respect to the position of the particle. Details about all the equation could be reported in an appendix to make the paper more readable. About the equation, personally I don't like the symbol composed of multiple letters such as Area, forcemag, forcedir, etc.

We have shortened the section as requested.

Do the authors use just one video for the training and two video for the testing? It seems that the dataset is quite small to have a systematic test of the algorithm. Even if there are only two objects, the authors should test different motions of the finger (i.e. pushing velocity, penetration, ...)

This is rather difficult to modify as we gathered the data set three years ago, and it would not be possible to extend it quickly enough using the same set up. However, to try to satisfy your request we are currently working on an extension of the data set using simulated data from a physics engine. While this will not be physically accurate it will allow further testing of the model. We are afraid we have not yet completed this work for this revision, as we focussed on completing the study of the comparable method, but we will submit our additional data and the results with a new revision when they are available.

I don't see the results of the classification. I would expect a confusion matrix to show how many times the sponge was classified as sponge and as plasticine and vice-versa. Table III seems to show the results about the classification of the pixels.

Vero?

Moreover, I would like to see a comparison against a baseline approach. They could use some of the models proposed in the literature or some simple models. For example, the authors could assume that the material "disappears" when the finger penetrates the object and report the F-Score of this method. Alternatively, they could use a simple model with only one spring between the finger and the base of the object.

We have implemented a growing neural gas network previously published by Cretu et al. We have added additional results in a comparison of this method with our own.

Vero?

Page 12. Col 2. Line 3. Please, can you clarify the hand tuning? It seems that the genetic algorithm is not the best way to find the parameters? Have you tried to use other method like gradient descent? Is it possible to use the model to mass-spring model to predict the force exerted on the finger? In that case, the authors could use this prediction error as fitness function and use the gradient descent to find the parameters.

Vero?

Other comments:

Fig. 1 is not referenced in the paper.

We have decided that this figure added little, hence it was not referenced naturally anywhere in the text. We have thus decided to remove it to save space, since other requested changes have added considerable length.

Page 2, Col 1, Line 17. Section VI is not mentioned.

We have fixed this.

Page 2. Col 2, Line 5: FEM is not defined. Finite-element models?

We have defined this.

Fig 2. Instead of use symbols, I would write Vision, strain, force, etc.

Thank you, this chimes with the comments of Reviewer 2, we have fixed this in Figures 1, 2 and 3.

Page 3, Col 1, line 57. Cretu et al. -> cite [10]?

We have added this reference.

Page 8. Col 2, Line 27. Linear snake = linear spline?

Vero?

Algorithm 5. I would specify that the F-Score is the fitness function of the algorithm. Moreover, it would be nice to know the values used for the parameters of the algorithm (perhaps using a table).

Page 10. Col 1. Line 45. Please specify that there are two objects, a sponge and plasticine, with a rectangular shape and so on.