Rethinking the Research Paper

By Frank Park

Like many of you, I've spent a good part of the summer months harried by paper deadlines. For each paper, the story unfolds something like this: the triumph and euphoria that one experiences immediately after submitting the paper quickly wears off, to be replaced by an encroaching regret that with more time, the paper could have been so much better. Then the reviews arrive: poring over the comments, decisions are quickly made on whether to accept or reject a criticism—should I hold firm or show some contrition for not explaining more clearly?—and whether to comply with all the requests (demands?) for additional experiments and comparisons to existing work. The road to publication is always a bumpy one, it seems.

While more evidence needs to be collected before any definitive conclusions can be drawn, experimenting with new review practices is a very welcome development.

Lately I've noticed that reviewers seem to be asking for more and more elaborate experimental comparisons against the state-of-the-art. At one level, it's a sign that robotics is maturing as a technical field. Our work is more and more embodied and validated by code, data, and benchmarks, and like the machine learning and data science research communities, there is growing acceptance within our community on the need to make code and data available.

What we still seem to lack in robotics, however, are benchmarks. To be fair, benchmarks have been developed for grasping, pick-and-place, and various other robotics tasks, but these tend to be narrowly defined for specific hardware and environment requirements and are difficult to implement. Simulation benchmarks have been proposed as an alternative, but even the best simulators today lack the ability to model friction, contact, deformations, and other complex physical interactions in a realistic way. I think it's fair to say that the inherent challenges of developing benchmarks for robotics are much harder than, say, those for vision or natural language.

Returning to the review of our paper, not surprisingly a reviewer had asked for comparisons of our method against another recently published method, claiming this to be the state-of-the-art. Since no code was provided for this state-of-the-art method, we spent a great deal of effort implementing the algorithm (whose description turned out to be lacking some small but crucial details) and performed the comparison experiments as faithfully as we could. We've not yet received feedback on these latest set of experiments, but already I can anticipate heated discussions on the experimental setup, algorithm implementation, and how truly meaningful these comparisons are in the absence of accessible, reproducible benchmarks.

I have also noticed that several our conferences are now experimenting with double-anonymous (or double-blind) reviews, open discussion phases between reviewers and authors, and several other new practices. All these efforts to try to reduce bias in our reviews—be it gender, nationality, author or institutional reputation, or any number of factors— are highly welcome and in keeping with our community's spirit of innovation and fairness. Some members of our community are currently studying the advantages and challenges of implementing double-anonymous reviews, as well as the experiences of other research communities. While more evidence needs to be collected before any definitive conclusions can be drawn, experimenting with new review practices is a very welcome development.

I encourage you to let your voices be heard in the ongoing community discussion on how we measure research progress, and exploring ways to further reduce bias in our review process.

Digital Object Identifier 10.1109/MRA.2022.3214390
Date of current version: 2 December 2022