Reviewer original text reported in italic.

REVIEWER 1

- I am inclined towards the original title "Better...".

ANSWER: changed to "Better vision through manipulation".

- Apart from (MAYBE) gaining a better model, the arguments for developmental approach need to be highlighted further. Such as, the building blocks idea - bottom up approach etc...

ANSWER:
The working hypothesis together with the description of the developmental aspects has been moved to section 1. We think this improves the paper by stating at the very beginning what the model is and what we're trying to demonstrate. We also clarified the description of table 1 and the idea of causation. This should make clear the developmental aspect of the discussion.

- The new introduction seem a little odd (especially the very first sentence), not all robots can observes, right?

ANSWER:
The word "observer" has been changed to "bystander". The point is simply that robots can take action, and experiment with the world. The choice of the word "observer" was unfortunate, since it implies vision, and as the reviewer rightly points out, not all robots have this sensory modality.

- Please provide a better presentation of the "three levels of causal complexity", numbering or using similar terms may help (e.g. 1) "sensorimotor coordination"), matching that of Table 1.

ANSWER:
Numbers have been added to table 1. Table 1 and 2 have been merged into a single table.

- The point made at the end of section 2, "The key to resolving ambiguity..." may be better serviced if highlighted/mentioned a little earlier.

ANSWER:
Unchanged. It's not really so fundamental having moved the "working hypothesis".

- Section 3 and 4 are nicely written.

ANSWER: Thanks.

- Please use a more up to date and accessible reference as well as Brooks et al. 1999. Such as Adams et. al. 2000/1.

ANSWER:
Added reference to Adams et al. 2000 In IEEE intelligent systems.

- I have completely fail to discovery where Table 2 was introduced?

ANSWER:
Thank you for the remark. Table 2 has been merged w/ table 1.

- Section 6, the stages are nicely presented. However, a better connection to biological processing (if any) need to be stated, otherwise it can be taken to be little misleading with the rest of the paper;

ANSWER:
A few links were provided in the section "A working hypothesis" now merged into section 1. In particular the reference to Flanders et al. and Pouget et al.

More details and links between section 6 and section 3 are now provided. Although this might help in clarifying the biological parallelism we don't think the experiments described in section 6 are "biologically plausible" in the traditional sense.

1. There's a link now between the detection of contingency between vision and motor commands that goes back to Piaget and it is also mentioned in the conclusion.
2. Localizing the arm using proprioception. We added a description and possibly linked it to the activation of VIP-F4 neurons that we described in section 3.
3. Reaching is implemented by building a mapping between "gaze direction" and "control". We linked this to section 1 where we mentioned the work of Flanders et al. and Pouget et al.

- the mapping discussed in "Reaching for the object", how does this mapping scheme account for ambiguity/redundancy (if any), in particular, for the case 3D mapping? Please discuss...

ANSWER:
We added a paragraph to explain the issue of redundancy. The mapping is not 3D but only 2D. Our goal was to reach an object sitting on the table.

- page 10: removing the 'a' may read better, "in this way to be a correctly segmented." -> "in this way to be correctly segmented." or, you need another word after 'segmented', "... a correctly segmented ??.";

ANSWER:
Fixed. Thanks.

- Is Section 9 really necessary for this paper? Without a ricer context it is difficult to see how this section fit in with the rest of the texts. Suggestion: shorten and/or save it for another paper.

ANSWER:
We think section 9 is important for the paper. We stated at the beginning of the paper that we wanted to show how something apparently complicated as "mirror neurons" can be the result of a developmental process based on much simpler hypotheses. We believe it represents an additional level of causal understanding (now numbered 3 in table 1).

- Overall this is a great paper, keep up the good work!

ANSWER: Many thanks to the anonymous reviewer.

REVIEWER 2

2. Are the research goals of this paper explicitly stated and is its terminology easily understandable by a heterogeneous community - including ethologists, computer scientists and roboticists?

Comments:

The goals could be made more clear. To the author's credit, in section 4, a "working hypothesis" is stated-- "action is required for object recognition in cases where an agent has to develop categorization autonomously". However, this hypothesis is a little lost in the other aspects of the introduction. First, a good deal of neuroscience is covered on pages 3-6 (it took me two readings to process this). While I do think the neuroscience adds to the theory aspect of the paper, this material comes before the "working hypothesis" and caused me some confusion about the goals of the paper.

GENERAL COMMENT:
We think the organization of the paper caused a bit of confusion and this is a good part of the issues raised by the reviewer.

ANSWER:
We moved the hypothesis and merged to section 1. We hope this clarifies and improves the organization of the paper.

The additional goal of our paper, perhaps a bit obscured in the text, is to:
"For the robotic implementation we endeavor to follow the same developmental pathway and exploit the same sort of causal links between actions and sensory feedback."

We added: "We wish to instantiate these results in robotic form to probe their technical advantages and to find any lacunae in existing models." to section 1 to make clear which are the reasons behind our experimenting.

Additionally, while the authors promise on page 1 to "discuss in what sense [their] artificial implementation is ... in agreement with neuroscience" there was no integration of the robotic work and neuroscience background in the conclusion/discussion section.

ANSWER:
There are now more links in section 6 (now 5).

Second, a series of developmental phases are proposed on page 6. Are these developmental phases being proposed for humans? Or for the nonhuman primates from which much of the neuroscience data comes?

ANSWER:
They are proposed for humans and/or for primates. We believed it was clear that we're not at the level of detail [of modeling] to distinguish between the two.

In either case, it would seem better to cite some ethological or psychological data regarding these developmental phases. Is there really a need to speculate about this level of developmental phases at this point in scientific history?

ANSWER:
We believe we had to show which model we were trying to implement. This is the reason of much of the neuroscience data and the description of developmental phases. Beside some of the work of Arbib (cited), there's not much of a firmly established model encompassing reaching, manipulation, and object recognition. We don't claim we are providing a very detailed model (in fact we didn't even try to do a precise localization of different modules/brain areas - many authors go to that level of explanation). What we wanted to show is how all the pieces fit under the same organizing principle (causation),

and roughly, functions/behaviors correspond to the activation of a set of brain areas (circuit).

The new organization of the introduction should make this point clearer.

We did cite some behavioral studies and little psychological. Having so many "direct" evidence from neurophysiology we didn't feel the necessity to include ethological considerations.

One major concern I have about this is the idea that grasping or at least hand motion comes first in their proposed developmental sequence. Perhaps this will be suitable for nonhuman primates, but the motor skills of humans may be relatively slower in development. Much benefit has come in developmental psychology by studying perceptual tasks.

ANSWER:
This is true and perhaps we should have paid more attention on alternative formulations. We were actually arguing in the opposite direction (and perhaps we've been a bit biased in this sense): that ACTION is very fundamental in case DEVELOPMENT and hence a natural setting is considered.

The neuroscience results we described extensively point towards an alternative view where perception becomes something more than just the "input of the system". In part perception is thus determined by action and in this sense it needs to follow from action (or sensorimotor coordination for some authors). Of course, all these consideration are very debatable - especially whether they're true in humans. Additionally we believe that one of the benefits of instantiating a model in a robot is that of validating whether this could be feasible.

This speculation on developmental phases confuses the hypotheses of the paper. Are these "developmental phases" really a description of the steps pursued in the robotic modeling?

ANSWER:
It is said now in section 1:
"For the robotic implementation we endeavor to follow the same developmental pathway and exploit the same sort of causal links between actions and sensory feedback."

If so, I think they should be stated as such. Additionally, the discussion and conclusion could more explicitly evaluate the working hypothesis.

Some of the terminology may pose difficulty for this audience. As I mentioned, it took me some time to work my way through the neuroscience. What might help here is to have a

little more focus on the working hypothesis, and then work this into the neuroscience section, perhaps putting the working hypothesis before the neuroscience section.

ANSWER:
The paper has been modified as suggested by the reviewer. The neuroscience section is a bit long, in our view, in order to describe the results and give a global picture to the reader. Perhaps, the way it's presented now, with causation as a guiding principle introduced right in section 1 should ease the reading.

Additionally, we replaced figure 2. Now all areas involved in a particular function are more clearly marked.

Reorganization of tables might also help.

3. Are the method(s) described in sufficient detail for readers to understand or replicate the work?

Comments:

> *For me, I?d need some more detail on how optic flow was computed and how the correlations were computed to be able to replicate this work. How about a reference to algorithms for each of the optic flow and correlation methods used?*

ANSWER:
Modified the description of the optic flow algorithm: "computed by a generic correlation-based approach over a 16×16 grid over a 128×128 image at 15 Hz"

"Correlation" was simply to be intended "cross-correlation".

4. Does the paper make contact with relevant earlier work in the same scientific area, noting similarities, differences and progress?

Comments:

There is some related work that could have been cited that was not. I just happened to come across the work of Scheier and Lambrinos (1996) recently. I think this could be profitably cited.

> *Scheier, C & Lambrinos, D.*
> *Categorization in a real-world agent using haptic exploration and active perception. In proc. of SAB96 (FROM ANIMALS TO ANIMATS, Fourth International Conference on Simulation of Adaptive Behavior, Cape Cod, Massachusetts, USA, September 9-13, 1996), p. 65-74*

ANSWER:
Cited.

6. Do its conclusions logically follow from the results obtained?

Comments:

As noted above, the working hypothesis, and neuroscience could be better integrated into the conclusions & discussion.

ANSWER:
Conclusions have been changed.

9. Is it well-organized and well-written?

Comments:

I think the organization could be beefed up considerably, and make what is already an interesting paper a better paper. This is notable in the experimental sections, which seem somewhat haphazard in organization at present. Some introductory material at relevant points would be valuable. Section 5 describes Cog, their robotic platform. It would be good to have a paragraph or two here in Section 5 that guides the reader through sections 6 through 10, the experimental sections on the basis of their hypothesis. For me, Sections 6 through 10 seem to flow too haphazardly at this point.

ANSWER:
A paragraph has been added as suggested.

Also, in section 6, the authors start off too quickly with optic flow. I would appreciate a brief consideration of alternative methods and/or at least their rationale for selecting optic flow as a method here.

ANSWER:
Added rationale - simply that optic flow allows us to make minimal prior assumptions about the appearance of the robot's arm. We don't think there really is a different approach if we want to make minimal assumptions about the appearance of the robot's arm, and just use the correlation between visual motion and commanded motion.

10. Additional comments:

The paragraph immediately before Section 4 seems incompletely integrated with the preceding material. For example, the proposal that there is a dorsal to ventral gradient of development seems interesting, and might be better tied into the preceding neuroscience. The idea of probing longer chains needs some elaboration in this paragraph as well.

ANSWER:
It was a rather speculative comment. We preferred to remove the paragraph because we cannot really support it appropriately with evidence, a part from looking at the localization of functions in the brain [see figure 2].

At the bottom of page 5, "three main conceptual functions" of what? Of causal understanding?

ANSWER:
Sentence changed into: “We can distinguish three main conceptual functions in the developmental process that leads to object representation (similar to the schema of Arbib et al. \cite{arbib-1981}).”

On page 6: "Learning and understanding affordances requires a slightly longer time frame since the initiation of an action (motor command) does not immediately elicit a sensory consequence". Does not immediately elicit ANY sensory consequence? Certainly some proprioceptive consequence will occur. I think what they are saying here is that some of the relevant sensory consequences will be delayed.

ANSWER:
Changed into:
"Learning and understanding affordances requires a slightly longer time frame since the initiation of an action (motor command) does not immediately elicit all relevant sensory consequences"

Is page 12 the first use of the term "active segmentation"? If so, perhaps it should be introduced earlier. It seems like a good term. Are the authors the first to use this term?

ANSWER: The term seems to be used in other contexts, particularly medical imaging. It could indeed be a useful term to import to active vision. It is now introduced at the end of the first section.

*Section 9?s use of a mimicry task seems to be trying to bite off too much for this paper. The task of object segmentation already seems pretty broad. I know the authors would like to try (as would *Many* others) to tie into the current fad of mirror neurons, but, for me at least, mimicry in addition to the main thrust of the working hypothesis of this paper bites off too much.*

ANSWER:
We think, this was another misunderstanding due to the organization of the paper. The mirror neuron experiment smoothly integrates with the other examples based on causation. We actually started a good part of our experiments with that in mind.

We think that with the reorganization of the introduction and initial sections of the paper this stands clearer. In fact, there's a full experiment (section 8 now) showing how mirror neurons can be seen as a derivation of "understanding object affordances".

The authors use the term ?causal understanding? without any citation of relevant literature. For example, this term is used at the top of page 6. Citation of work such as:

> *Dan Sperber, David Premack, and Ann James Premack (eds.)*
> *Causal cognition: A multidisciplinary debate*
> *New York, NY: Oxford University Press, 1995*

would be suitable here.

ANSWER:
The reference has been added. We were relying on the intuitive meaning of causation but of course it's much better to have a proper link. Thanks to the reviewer.

The organization of the introduction regarding hypotheses should be improved as should the organization of the experimental sections. The discussion should be revised as mentioned.

I think this is a very interesting line of research, and this paper has the potential to be very good. I would be happy to comment on the paper again prior to inclusion in the journal.

ANSWER:
These were general considerations at the end of the review. We believe they are automatically answered by the changes we described above.

REVIEWER 3 [PARTIAL REVIEW]

I read pp. 3-5 of Meta and Fitzpatrick. Mostly, it's a reasonable review of recent work on cortical systems involved in reaching, grasping, and object recognition. However, in the first part of this section, they've made it clear that they do not understand the different ideas of Ungerleider and Mishkin vs. Goodale and Milner on the functions of parietal

and temporal cortex. They attribute the dorsal/ventral distinction to Goodale and Milner, when actually it should be attributed to Ungerleider and Mishkin. Also, they attribute the idea that the dorsal pathway is about "vision for action" to Ungerleider and Mishkin, whereas the idea of "vision for action" was developed by Goodale and Milner in response to what they saw as a deficiency in U & M's account of the functions of the dorsal pathway. Recall that U & M characterized the functional differences between the ventral and dorsal pathways as matters of "what" and "where."

ANSWER:
We made a mistake, references and description of the work of U&M vs. G&M was wrong. We fixed the issue. Many thanks to the anonymous reviewer.

REVIEWER 4

2. Are the research goals of this paper explicitly stated and is its terminology easily understandable by a heterogeneous community - including ethologists, computer scientists and roboticists?

Comments:

The goals are reasonably stated, but not very clear, especially when they are swamped under an introduction that is largely of a tutorial nature. I suggest that the article clearly state exactly what has been done in the abstract and in the introduction (with more detail).

ANSWER:
We changed substantially the introduction (see also reviewer 2) so we believe that now goals and scope of the paper are more reasonably stated at the very beginning.

3. Are the method(s) described in sufficient detail for readers to understand or replicate the work?

Comments:

The work presented is a good example to show how manipulation helps segmentation. Since this is an early work, we should not require it to work in an uncontrolled setting. However, for the understanding need, I suggest that the paper spell out all the major restrictions, about the environment and the task(s).

ANSWER:
A summary of the restrictions of the work has been added in the conclusions.

4. Does the paper make contact with relevant earlier work in the same scientific area, noting similarities, differences and progress?

Comments:

I like the work presented here, since it argues convincingly and demonstrates experimentally the importance of manipulation in infant cognitive development. However, the paper should cite the line of recent studies on autonomous mental development (AMD) that approaches the principles addressed here in a systematic and autonomous fashion, such as the work of Gerald Edelman and Olaf Sprons (manipulation of objects to learn its properties, "aversive" or "appetitive") and John Weng (a human teacher manipulates objects for autonomous attention and object recognition).

ANSWER:
We added references in the general discussion at the end of the paper. “Many researchers have shown now examples of the application of developmental principles in the design of autonomous systems, for example \cite{weng-2000, weng-2002} and \cite{metta99developmental}. This approach may provide novel directions to robotics. Besides, it may also serve as a useful reference point from which to investigate the biological solution to the same problem -- although it can't provide the answers, it can at least suggest useful questions.”

5. Does it make contact with relevant work in other scientific areas?

Comments:

The work is motivated well. I just feel that the material of discussion about biological brain (much of Section 3) is of a general tutorial nature and it does not link the work presented here specifically. Any paper about perception and/or manipulation can come up with this tutorial section. I suggest that the material be dropped or much condensed into a brief subsection if some is truly intimately related. The material is fine for a PhD thesis, but not appropriate for a journal article.

ANSWER:
We added more links from the presentation of the experimental sections to the biological part that make the neuroscience results more relevant. It was intended as a support to our theory rather than just as a tutorial. It is for the benefit of a general audience although we agree the paper could have been written differently.

What is lacking is a discussion about how manipulation is important to early cognitive development in infants and young children and in what ways. Such literature is abundant in developmental psychology.

ANSWER:

Added a brief comment and citation in the introduction. While this literature is very relevant and interesting, to do it justice would further increase the density of biological background in the paper, which another reviewer felt was already burdensome.

It is also useful to add a brief discussion about a school of psychology which postulates that human children were born with some knowledge about the physical world (e.g., Baillargeon, Spelke and others). The last section alluded to it.

ANSWER:
See previous answer.

6. Do its conclusions logically follow from the results obtained?

Comments:

It is good to discuss the cost of such an approach compared with traditional human trial-and-error approach from engineering point of view. In a controlled setting, there have been many studies about object segmentation, which use similar ideas (e.g., motion based segmentation) as what are presented here. A major difference between those past studies and what is presented here is that here it is the robot who initiates the manipulation, which is not necessarily a part of the originally given task. Then, there is a cost issue.

ANSWER:
Indeed, in our approach information and training data are not for free. There's a cost associated to collecting information. The same, we believe, it is true for any system which is not based on receiving explicit training (as in the supervised learning case). One nice (and similar) example where this dichotomy of acquiring new data vs. accomplishing a task is presented is "reinforcement learning" (e.g. see Sutton & Barto, RL: an introduction). In the reinforcement learning approach there's always a tradeoff between exploration of the state space and "exploitation" of what has been learnt up to that moment in solving a task.

In our approach, poking is only needed when interacting with an object for the first time. And it can be used to learn autonomously about objects and their visual appearance. Afterwards, poking can be effectively inhibited (as a behavior) and object can be manipulated without relying in this relatively costly operation.

We mention this aspect at the beginning of section 8 (now 7): “This kind of active segmentation will nevertheless be inconvenient in many situations if not coupled with a mechanism to learn from experience”. That is, there's a cost issue but this is a price to be paid only during development.

7. Does the paper consider the implications of the approach and outline directions for future work?

Comments:

If it is the programmer to program such manipulations, then the current work and the traditional ones fall into the same category in this respect (which is fine). However, in other to attack the core of the subject raised by this paper, we must address how to enable a robot to autonomously initiate such manipulations in uncontrolled settings. The paper could raise this future research issue.

ANSWER:
It is the robot that decides when to poke an object. The rationale is as follow: the attentional system can certainly DETECT (but NOT segment) objects on the basis of a pure bottom-up saliency mechanism. A reaching movement can then be initiated. If the object is touched and a "good" motion signature is available the segmentation algorithm might produce a segmentation. To initiate this procedure the robot doesn't need any "programmer" input beside that of moving objects close enough to the robot.

Although limitations (see also the conclusions) are definitely present the environment is relatively uncontrolled. If the object is missed (poking fails) or the signature is poor then the data is discarded. Statistically, if the robot "plays" with an object long enough then a simple object model (section 8 - now 7) is constructed.