3D vision: subjunctives

D.A. Forsyth

OR: “Go to the ant…; consider its ways and be wise”

What does vision do? (traditional)

- Recognition
  - instances (who is this?)
  - allocate pictures of objects to categories (classification)
  - find location of objects in pictures (detection)
  - produce descriptions of objects (attributes/primitives)
  - describe pictures (captioning)
- Reconstruction
  - SLAM
  - point clouds
  - meshes
  - voxel reconstructions
  - geometric primitives
  - implicit surfaces, generalized cylinders, superquadrics, etc.
- Lots of evidence these threads interact
- Lots of evidence that these activities have created value

What are we really good at?

- Classification
  - eg image classification; voxel labelling; detection (== lots of classification)
  - in the presence of huge quantities of labelled data
- Regression
  - eg predicting boxes; depth; voxels; etc.
  - in the presence of huge quantities of labelled data
- (Some kinds of) Geometric reasoning
  - SFM writ large
- Our actions are driven by our tools (OADOT)

What are we bad at?

- (Almost) Unsupervised learning of visual representations
- Controlling the bias of representations for advantage
- Will reinforcement learning save us?
  - NO

OADOT - Recognition

- Categories clearly don't exist in any canonical sense
  - and any instance can belong to many different categories, etc.
    - be very careful of:
      - members of a category share properties or are alike
      - what properties? in what sense alike?
- And so *MUST* be the product of unsupervised learning
- Categories are useful intermediaries
  - it is helpful to group instances together in clusters that
    - improve prediction
      - dog-a will very likely behave in the same way as dog-b
    - improve communication
      - it's easier to talk about dogs than dog-a, dog-b

OADOT - Reconstruction

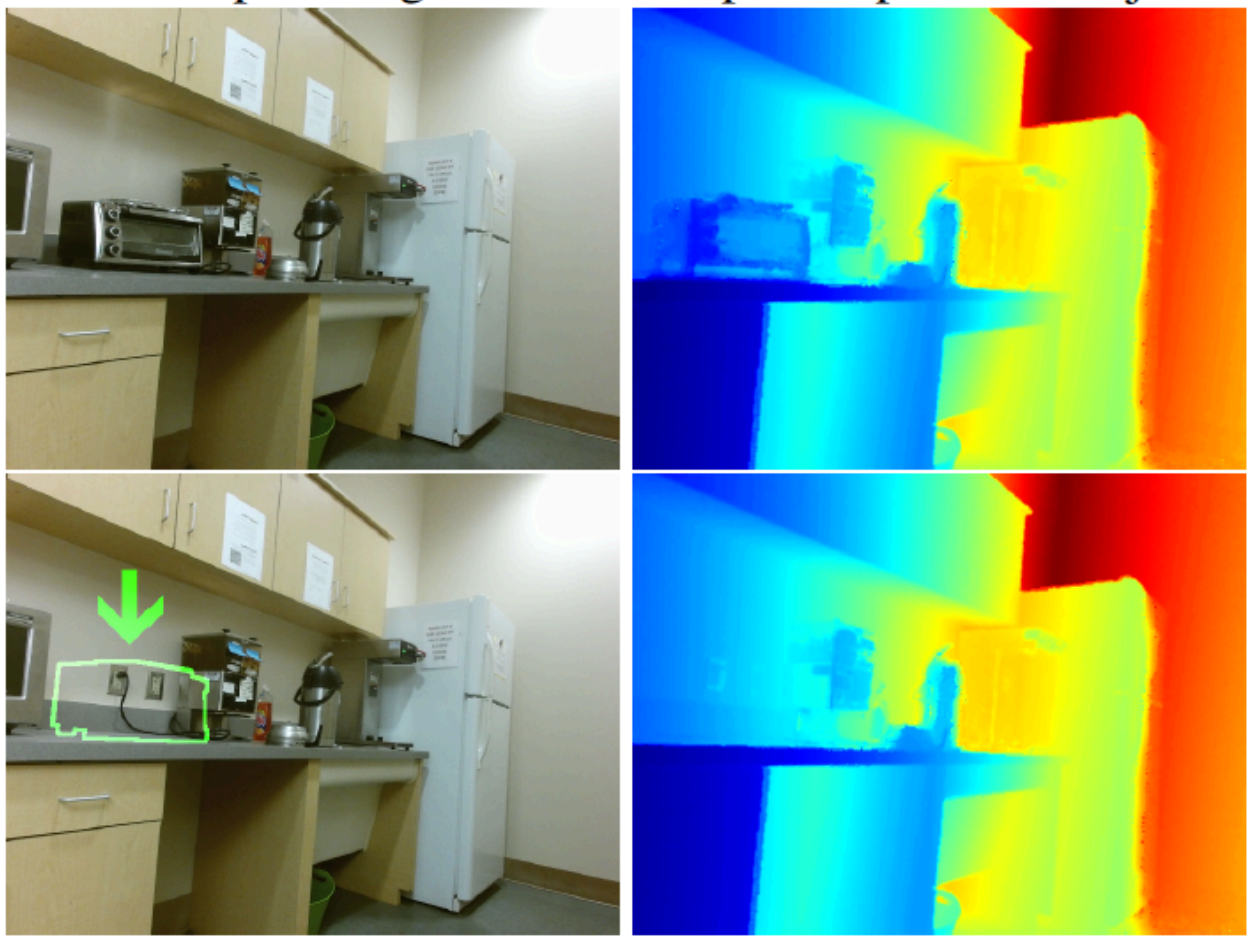
- Reconstructions don't exist in any canonical sense, either
  - there really isn't any single 3D representation cause there can't be
  - there is no evidence that *any* visual task *requires* a 3D rep'n
    - Q: how can you determine *from outside* whether an agent has one?
- 3D representations are intermediaries
  - and useful to the extent they mediate
    - eg: point clouds, meshes
      - renderable models; metric info; maps
- What task does this representation facilitate?
  - what info does the task need?

What problems to focus on?

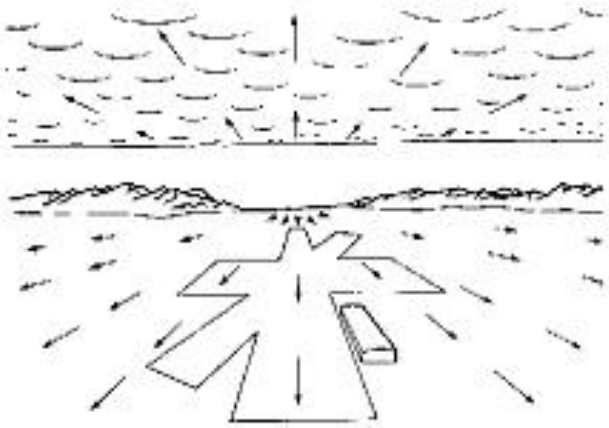
- Improved geometric models from images is always good
  - there's a reason to care, etc.
- Orphan problems
  - The space we can't see
  - How do I know there is a 3D world?
- Functional problems
  - Where am I?
  - How do I get home?
  - What could I do?
  - What might happen?

The space we can't see

- Speculated depth
  - what would depth map look like if an object was removed?
  - what is behind closest object?
  - could I move there?



How do I know there is a 3D world?



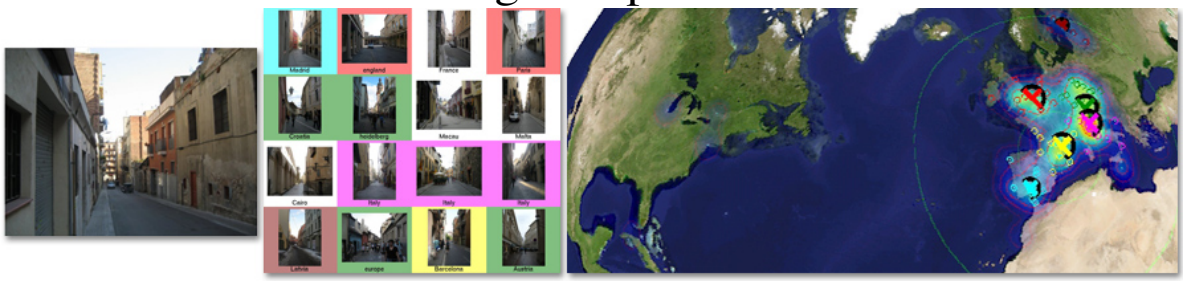
- and how to act in it?
  - (without invoking RL)

- Various answers:
  - 3D means textures are more uniform (Fouhey et al 15)
  - the parametric forms of flow fields are more easily explained (Gibson, 50)

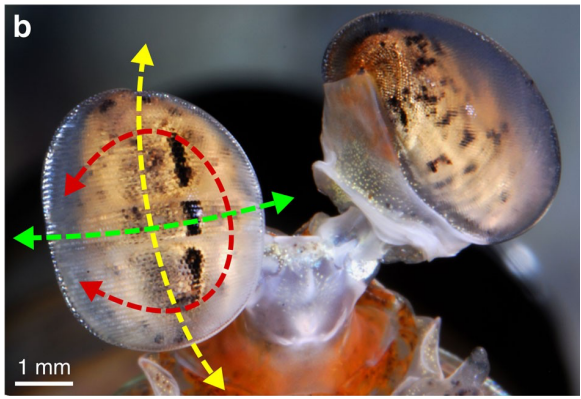
- Do I need to know there is a 3D world?

Where am I?

- This doesn't get sufficient credit as 3D
  - early work (im2gps, etc; Hays+Efros 2008)
  - non-par regression (matching)
  - NOT the same as building a map



- Short scales, visually simple worlds are hard
  - get different visual sensors and use them well
    - Mantis shrimp (Daly et al 2016)



- Movie

How do I get home?

- Desert ants can forage, then go home directly
  - They're not doing SLAM! (scale)
  - Cues:
    - dead reckoning (count leg movements)
    - visual waypoints
    - polarization based sun compass
- Behavior can be explained *without* a map
  - multiple cues each produce a "go-home" vector
  - weighted combination (Hoinville+Wehner, 2018)
  - can be imitated (Dupeyroux et al 2019)
- And they can go home backwards

- Movie

What can I do?

- Path planning is not about geometric detail
  - which creates computational complexity
  - RRT methods; nearest neighbor methods; = strategies to duck detail
    - the key is a test: will this result in collision?
  - So why recover detail from images, rather than be able to answer query?
- We should recover geometric affordances of objects
  - what can be done to this, and where?
  - this likely isn't inherited from category
- Does a clam shell have a "hit here" tag?

- Movie

What might happen?

Conclusions

- What we do is shaped too much by our tools
  - collect dataset - regress - repeat
- 3D representations are mostly intermediaries
  - the ones we use should be task appropriate, not generic
- Appealing problems:
  - The space we can't see
  - How do I know there is a 3D world?
  - Where am I?
  - How do I get home?
  - What could I do?
  - What might happen?

Structure

- traditional view:
  - recognition
    - instance: - useful for some special cases
    - categories: - clearly don't exist in any canonical sense, but are very useful intermediaries
  - reconstruction
    - various geometric representations: - typically intermediaries
  - lots of evidence of interaction

- What can we do?
  - regression
  - classification
  - both really well, in the presence of large, labelled datasets

Structure

- what should vision do?
  - inform action
    - pure reinforcement learning is ridiculous, so representations are needed
  - what to recover?
    - current geometric representations are inconvenient devices
  - perhaps
    - break out representations by the problems they can be used to solve
      - exploration
      - going home
      - interaction
      - prediction