CS 280 Computer Vision
Spring 2018 Efros, Malik, Yu Lecture 1

1 Lecture

The purpose of today's class is to go through a whirlwind tour of where this field is today.

The Three R's

In 1998, Moravec estimated the computing power of the human brain and compared it with computers of the time. Apparently, he found that we would obtain human-level computing power in 2025 assuming Moore's law. We could also obtain it with a cluster of 100 nodes in 2015.

Vision does not exist in isolation; it is connected deeply to motor control (planning) and language/semantic reasoning. 543-490 million years ago the first multicellular animals with vision and locomotion capabilities appeared, and this is where vision began. Note that sensory systems provide great benefits only when accompanied by the ability to move (to find food, avoid predators, etc.). Hence the joint evolution of vision and locomotion! The bottom line is that vision is for a purpose, and one of those purposes is locomotion. Another is manipulation.

Evolutionary progression: vision and locomotion, then manipulation, then language. Successes in AI seem to follow the same order! Note: when it comes to manipulation and AI, hands are a limitation right now. Robots typically have only two-finger grippers, whereas humans have five-fingered hands with high dexterity. There is no such limitation with vision: all we need is a camera, so we are not held up by our sensing device in the same way that robots are held up by their actuating device(s). Perhaps as a result, vision and locomotion are the most advanced of the subfields listed earlier.

The three R's are **recognition**, **reconstruction**, and **reorganization**. These are all different aspects of the vision problem, and represent how we structure the field.

- **Recognition:** attaching category labels to objects that we see... or, more broadly, *connecting individual instances that we see to examples that we have encountered in the past.*
- **Reconstruction:** reconstructing the 3D world from which this 2D world arose (the inverse of computer graphics). For example, what is the depth? What is the illuminance? What are the surface normals? Everything a graphics person needs in order to produce pixels is what we want to determine now.
- **Reorganization.** An image appears to a computer as a set of pixels, but it ultimately consists of objects. Going from a representation of 1000 x 1000 pixels to these objects is what psychologists call **perceptual organization** (or **grouping**, or **segmentation** – where we start with the full thing and break it into separate entities).

There are bidirectional arrows between each of the three R's, signifying that any of these paths is a valid and useful direction of information flow. We do not necessarily perform any of the three R's before each other (e.g. reconstruction before recognition). Instead, they are all interconnected processes which are happening simultaneously, with information flowing back and forth among themselves.

Accordingly, we can use one R to help with another. For example, recognition can help with reconstruction (see Category-Specific Object Reconstruction). In another example, reorganization/grouping can help with recognition: even without any experiential knowledge of specific object categories, we can produce **superpixels** (groups of pixels with the same property, such as color or texture). Then we can combine neighboring superpixels to generate object candidates.

RGB values, where each pixel equates to three numbers, are what we *literally* have in a picture. That's what the stupid computer sees. But what do we, as humans, see?... a very rich scene with human people and objects. Our goal in computer vision is to get to all of that (to be able to make the inferences of which humans are capable). *How do we get from pixels to these inferences?* That is the million-dollar question.

A Brief History of Computer Vision

- **1960s:** the beginning of computer vision – with AI, image processing, and pattern recognition.
- **1970s:** basic math of image formation.
- **1980s:** vision became its own field (separate from general AI). Essentially, it became applied math.
- **1990s:** finished geometric analysis, started using statistical learning approaches again.
- **2000s:** major advances in visual recognition.
- **2010s:** large amounts of visual data, massive computing power (before, this didn't exist; we used to have all the *lack* of computing resources that a backward-facing Moore's law suggests). Deep learning runs rampant.

Phases of Computer Vision – Geometrical Aspects

Multi-View (***how do we know something about the 3D world from multiple images?***)

- **Binocular stereopsis:** see a projection of the world in the left and the right eye. The images will be slightly different – this difference is the binocular disparity, which is a signal from which we can estimate depth. Disparity is a signal that is only good near to you, because it falls off as $1/z^2$ (i.e. when something is very far off, the disparity is practically 0).
- **Optical flow:** a continuous disparity signal corresponding to a single moving camera (*relative motion*). This is a basic cue for all animals. The apparent movement of a pixel, corresponding to some feature on the ground as a plane takes off, constitutes optical flow.

Say we have a left and a right camera. Between the two cameras, say there is some rotation and translation. If I see a point in one view, it must lie somewhere on the line from that camera's center of projection. Then, if I see the same point in the other view, it must lie somewhere on the line going out to infinity through the *other* camera's center of projection. Thus our point, in 3D space, must lie on the intersection of these two lines (since it has to lie on both of them). Simple, right?

Not quite. Unfortunately, we are unlikely to know the rotation or translation between the two cameras. This makes things a little more difficult.

Single-View

We can see depth even in a single image – we don't need two. As humans, we are able to take pictorial cues and infer depth and shape as a result of our experiences with the world (which has certain regularities, such as light usually coming in from above). *This is a matter of experience, and should therefore be learnable.*

Successes in Computer Vision

Reconstructing the World

We can construct scenes automatically from huge collections of photos (e.g. from tourists uploading photos from the web). To accomplish this, we can download all pictures of the Colosseum of Rome, which will be taken from different perspectives, and build beautiful reconstructions of it and whole cities besides.

We can reconstruct indoor scenes from the data captured by moving around with a depth camera. This is important for the businesses of AR and VR (which need the real world to be captured within a computer and therefore rely on techniques such as this).

We can bring in semantics. Imagine reconstructing a room; we want to know that some set of pixels represents a table and thereby a group. We should be able to, for example, place objects on such a table.

Convolutional Neural Networks

The pattern of weights is replicated across space.

R-CNN: Regions with CNN Features ("r" stands for "region")

In one image, we have many objects. We want to find all of them. Thus we need to isolate the pixels of each object, which we can do with region proposals (e.g. from superpixels). We can then feed each of *those* sets of pixels to a classifier, and end up with an object category for each one.

Current systems are remarkably good at detecting objects.

Mask R-CNN: don't just recognize an object, also mask its pixels! This is the current state of the art.

Takeaways

Vision connects to cognition and understanding. The goal is to get to visual understanding.

We would like computers and humans to work together. However, currently computers have low social intelligence and can't understand the internal state of humans very well (e.g. emotional state, body language, current goals). So another aim is to create good vision techniques for understanding the hierarchical structure of human behaviors (movement, goals, actions, events; note that *action = movement + goal*).

In general, we want to build models of activities and behavior. An organism should possess a small-scale model of external reality and its own possible actions, so that it is able to make better decisions. A common computer vision aphorism is that "parts assume their significance in the context of the whole." The same is true for individual actions: the significance of any two seconds of the action is in the context of the bigger activity. And to acquire that context we need models, which are learned from data.

How to acquire that data? We can either use external supervision or self-supervision. In the past, we have fed models tons and tons of external teacher signals. However, this is not ideal (after all, human children do not need to be told "this is a zebra" 10,000 times). We should have a *continually growing model.* What we'd like to do is simulate a child's brain (an idea dating back to Turing in 1950) and manage it with internal teacher signals. According to Prof. Malik, this is the paradigm to explore right now.

Finally, there are many lessons that the fields of computer vision and robotics can learn from studying the course of human development (i.e. what children do):

- **Be multi-modal.** Children are multi-modal; they have access to sound and touch and their internal signal all at once.
- **Be incremental.** You don't start teaching calculus to a six-month old baby. There's a sequence in which things happen.
- **Be physical.** Children interact with the world.
- **Explore.** Children perform lots of experiments.
- **Be social.** Children acquire experience from teachers and caregivers; from them comes an accumulation of knowledge thousands and thousands of years in the making.
- **Use language.** This adds a symbolic aspect to learning, as well as a means of communication.