Lecture 10
An Introduction to Computer Vision: Part II

Rob Gaizauskas

Lecture Outline

- What is Computer Vision?
- Image Formation
- Early Image Processing Operations
- Object Recognition by Appearance ⟵ Today
- Reconstructing the 3D World
- Object Recognition from Structural Information
- Applications

- Reading: (Readings that begin with * are **mandatory**)
  - *Russell and Norvig (2010), Chapter 24 “Perception”

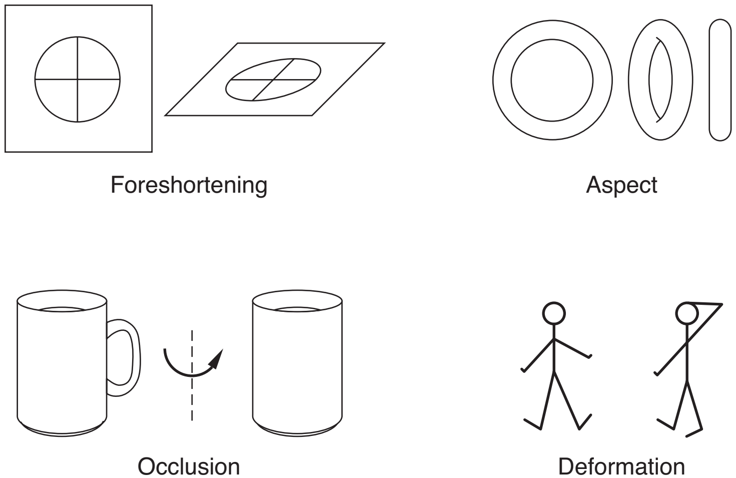
Object Recognition by Appearance

- Some object classes vary little in appearance. E.g.:
  - soccer balls; bricks
- Other object classes vary a lot. E.g.
  - houses: vary in size, colour, shape and look different from different angles
  - ballet dancers: look different in each pose, under different lighting

Object Recognition by Appearance

- Can test images for objects of specific categories using a classifier trained on image data with objects of the category marked within a bounding box
- Recognition method:
  - To find objects at different scale/location in the image: sweep windows of varying size across the image, testing each window with classifier
  - Can extend to consider different orientations too …
- Classifier approach
  - Works well for e.g. faces – most faces look quite similar – round face, face bright compared to eye sockets, mouth and eyebrows dark slashes
  - Doesn't work so well for more variable object classes

Object Classification



- Effects that cause problems for object classification include:
  - Foreshortening
  - Aspect
  - Occlusion
  - Deformation

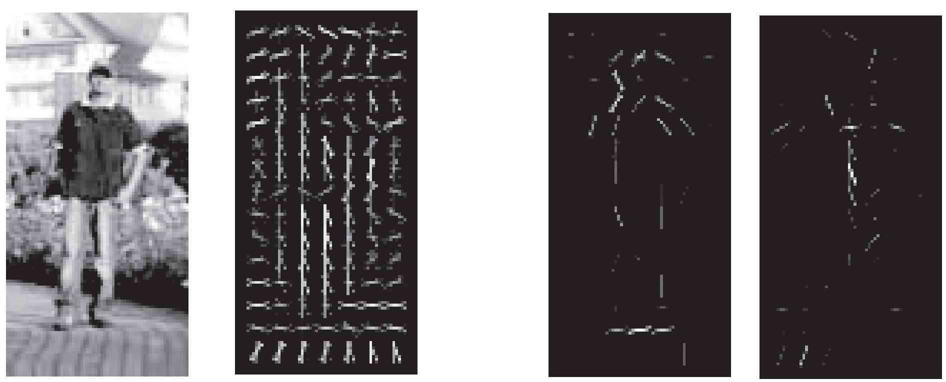
Object Classification

- Classification approach can still work, but may need to represent complex objects as collections of parts
  - E.g. car image likely to show some of headlights, doors, wheels, etc. in similar arrangements
- So, model objects with pattern elements – collections of parts
  - Pattern elements may move around wrt each other
  - If most elements are present in about the right place, then the object is present
- Object recognizer then a collection of features testing for presence of pattern elements and their location in about the right position
  - Can use, e.g. histograms of pattern elements + some spatial info

Object Classification

- Until recently two feature-based approaches for object classification obtained best results:
  - SIFT: Scale-invariant feature transform

  - HOG: Histogram of gradients



Image

Orientation histograms

Positive components

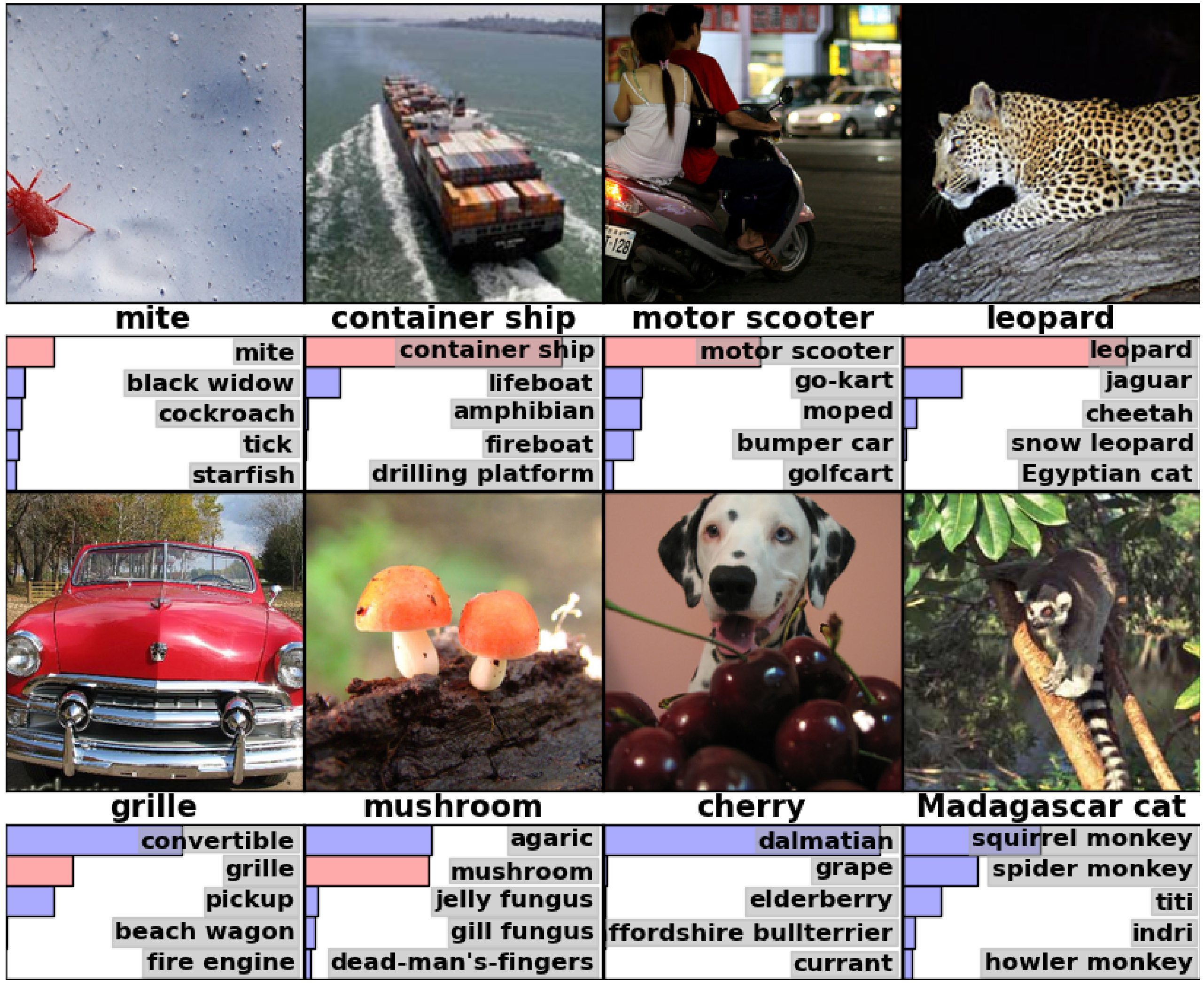
Negative components

Object Classification

- State-of-the-art results now obtained using "deep learning"
  - Specifically, convolutional neural networks (CNNs)
  - CNN: "a type of feed-forward artificial neural network where the individual neurons are tiled in such a way that they respond to overlapping regions in the visual field" (Wikipedia, "Convolutional Neural Nets")
- ImageNet – dataset with ~15 million high res images in ~22,000 categories
  - Annual challenge: ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)
  - uses ~1000 images in each of ~1000 categories
  - 1.2 million training images, 50,000 validation images, and 150,000 testing images.
- CNN has best results:
  - top-1 and top-5 error rates of 37.5% and 17.0% on ILSVRC 2010 dataset – considerably better than any previous approach

Object Classification

- Example test images from ILSVRC-2010
- Correct label under image
- Top five hypotheses from Krizhevsky et al. 2012
  - Bar indicates probability
  - Red bar correct



Lecture Outline

- Object Recognition by Appearance
- Reconstructing the 3D World
- Object Recognition from Structural Information
- Applications

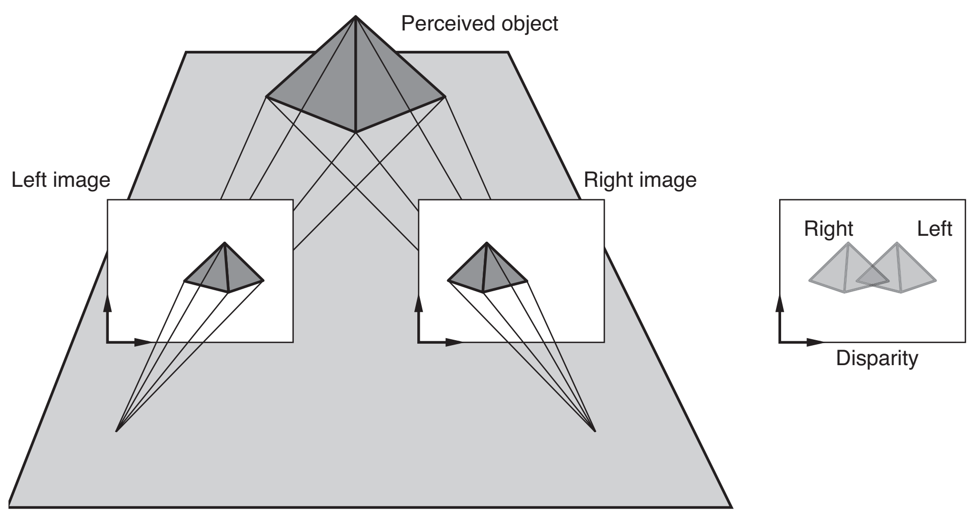
- Reading:
  - Russell and Norvig (2010), Chapter 24 “Perception”

Reconstructing the 3D World

- How can we recreate a 3D representation from a 2D image?
- Cues that can be exploited include:
  - Motion (of camera or of objects in scene)
  - Binocular stereopsis/multiple images from different camera positions to triangulate and find depths of points in image
  - Texture and shading
  - Contour/background knowledge about familiar objects or physical scene types that gave rise to the image

Cues for Scene Reconstruction

- Motion
  - When a camera moves relative to a 3D scene optical flow can reveal relative depth of objects in the scene
- Binocular stereopsis
  - Most vertebrates have two eyes
  - Multiple images separated in space let us use the disparity between the images to compute depth – but must solve correspondence problem
    - determine for point in left image the point in right image that results from projection of the same scene point)



(a) (b)

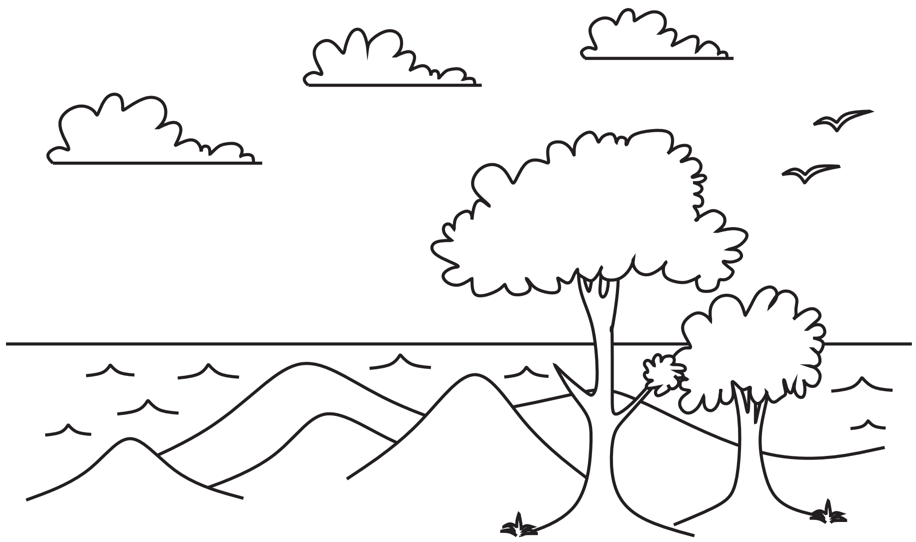
Cues for Scene Reconstruction

- Multiple views
  - Can use more than 2 cameras
  - Need to solve
    - Correspondence problem – identifying features in different images that are projections of same feature in 3D world
    - Relative orientation problem – determining the transformation (rotation/translation) between coordinate systems fixed to different cameras
    - Depth estimation problem – determining depths of points in the world for which image plane projections were available in at least two views

Cues for Scene Reconstruction

- Texture
  - Can use texture to estimate distance
  - Homogenous texture, e.g. bricks, tiles, paving stones, results in texture elements or texels in the image
  - Real world elements giving rise to texels are typically identical; differences between texels in image due to
    - Distance – more distant elements appear smaller in the image
    - Foreshortening – more distant elements in the ground plane are viewed at an angle farther from perpendicular and so are more foreshortened
  - Magnitude of these effects can be used to calculate distances
- Shading
  - Variation in light intensity is determined by geometry of scene and reflectance properties of surfaces – can try to use image brightness to recover geometry + reflectance – hard!

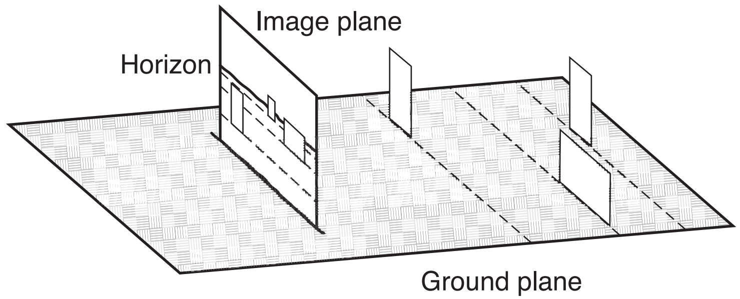
Cues for Scene Reconstruction



- Contour
  - How do we infer 3D shape from simple line drawings like this?
  - Combination of recognition of familiar objects plus generic constraints such as
    - Occluding contours, such as hills: one side of contour is nearer, other farther away – can use local convexity/symmetry to solve figure/ground problem
    - T-junctions: when on object occludes another, the contour of the farther object is interrupted and a T-junction results
    - Position on ground plane: humans and other object typically on ground plane (due to gravity) – can exploit this -- bottoms of nearer objects project to points lower in the image plane; farther objects have bottoms closer to the horizon

Cues for Scene Reconstruction

- Familiar objects



  - E.g. Can use knowledge that people
    - Have heads that are about 9 inches
    - Are about the same size
    - Tend to stand on the ground

  to do relative positioning of pedestrians in a street scene

Lecture Outline

- Object Recognition by Appearance
- Reconstructing the 3D World
- Object Recognition from Structural Information
- Applications

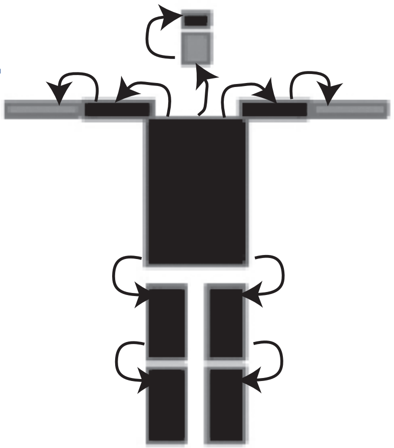
- Reading:
  - Russell and Norvig (2010), Chapter 24 "Perception"

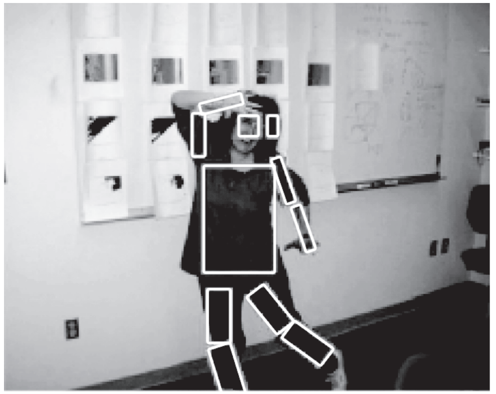
Object Recognition from Structural Information

- For objects that are highly deformable, it may be necessary to
  - Recognize them in different configurations
  - Determine the disposition of their constituent parts
- For example, human bodies:
  - May need to identify them in many different poses
  - May want to know disposition of limbs to determine what activity person is engaged in (e.g. running vs sitting)
- Can use a model called a deformable template model to specify acceptable configurations
  - E.g. elbow can bend, but head is never connected to foot

Geometry of Bodies

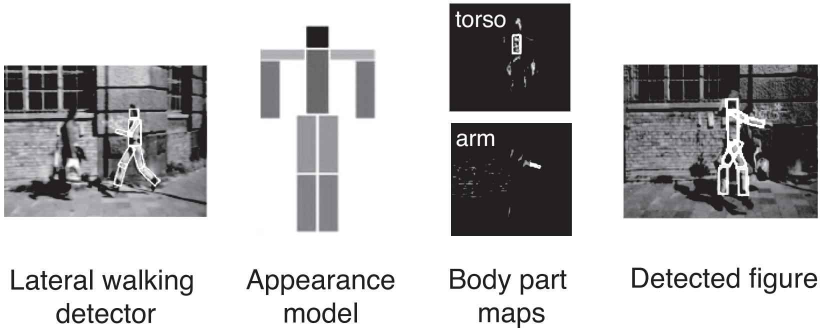
- Can model body as tree of 11 segments: upper + lower right + left arms + legs, torso, face and hair
  - Called cardboard people models

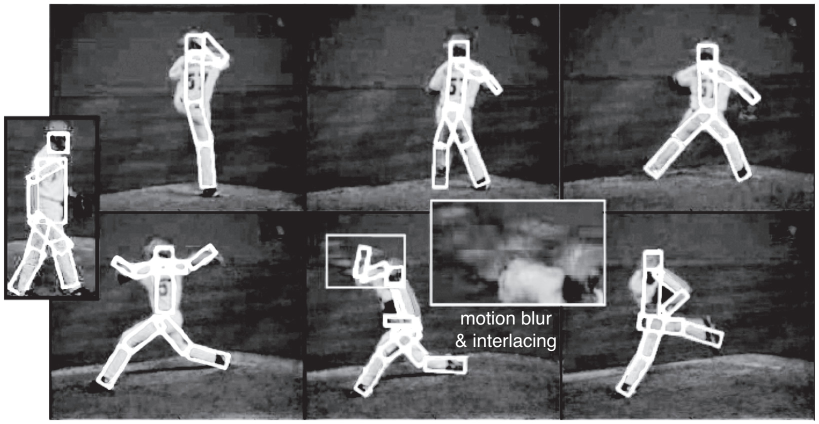




- Then search image for best match to cardboard person using probabilistic inference methods. Two main matching criteria
  - Image rectangle should look like its segment
  - For each pair of related body segments should be a pair of image rectangles with relations that match those of the body segments
- Can use pictorial structure model to compute matching scores of different candidate image rectangle configurations

Tracking People in Video





- Track people by using pictorial structure model to detect person in each frame
- Get pictorial structure model by first obtaining an appearance model
  - Models what segments look like
- Obtain appearance model by
  - First scanning to find lateral walking pose
  - Find pixels that lie in each body segment and those that do not
  - Develop discriminative model for each body segment from this
- Pictorial structure model then obtained by using appearance model + constraints on segment connections

Lecture Outline

- Object Recognition by Appearance
- Reconstructing the 3D World
- Object Recognition from Structural Information
- Applications

- Reading:
  – Russell and Norvig (2010), Chapter 24 "Perception"

Applications

- Vision systems that could analyze video and understand what people are doing could be used to
  - Inform better building/urban space design by collecting and analysing data about how people use them
  - Build better person tracking for surveillance
  - Gather information for military (or other) intelligence: DARPA's Visual Media Reasoning Concept Video
  - Build computer sports commentators
  - Build reactive interfaces for computer games, energy saving systems in buildings
- Autonomous vehicles
  - Autonomous driving: DARPA Urban Challenge
  - Autonomous submersibles, unmanned aerial vehicles (UAVs), for, e.g. forest fire detection, crop monitoring

Applications

- Automatic Image Captioning
  - Can we generate textual image descriptions?
    - Useful for indexing, searching and retrieving images
  - VisualSense project's use of SoA classifiers at large scale
- Video Search
  - Find recent news clips containing: David Cameron, bicycle races, advertisements

References

Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton (2012). ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012: 1106-1114.

Snowden, Robert, Thompson, Peter and Troschianko, Tom (2012) Basic Vision: An Introduction to Visual Perception (revised edition). Oxford.

Szeliski, Richard (2011). Computer Vision: Algorithms and Applications. Springer.

Russell, Stuart and Norvig, Peter (2010) Artificial Intelligence: A Modern Introduction (3rd ed). Pearson. Chapter 24.

Wikipedia: Computer Vision. http://en.wikipedia.org/wiki/Computer_vision (visited 01/11/15).