Computer Vision

COMP9414: Artificial Intelligence

Lecture Overview

- Introduction
- Image processing
- Scene analysis
- Cognitive vision

Lecture Overview

- Introduction
- Image processing
- Scene analysis
- Cognitive vision

Introduction

- Other sensory modalities an agent uses for interaction with the world (vision, acoustic, temperature, pressure, etc.)
- Computer vision endows machines to “see” the world.
- Applications in character recognition, image interpretation, face recognition, fingerprint identification, robot control.

Introduction

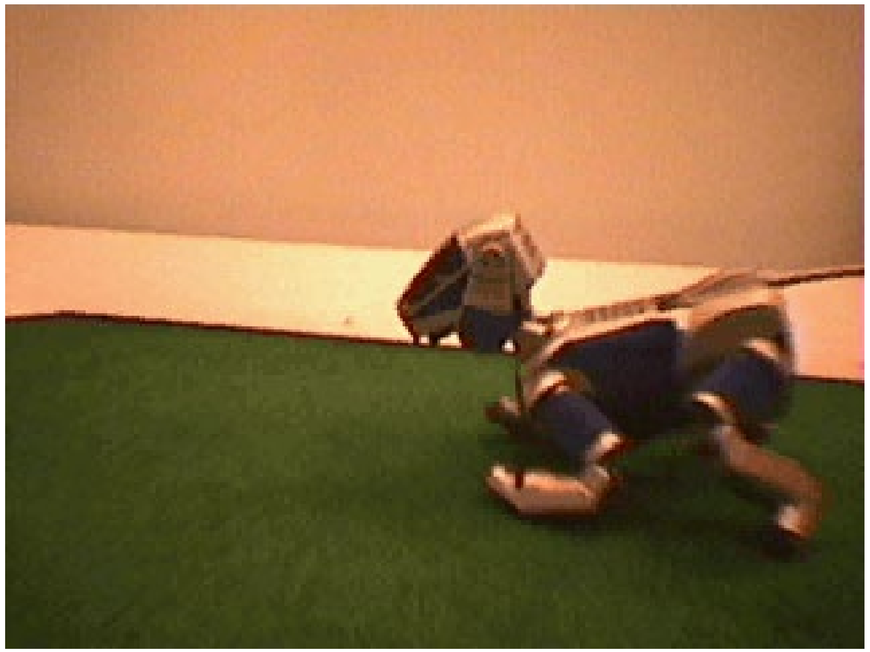
- Effortless for humans, but a difficult problem for machines:
  - Variable and uncontrolled illumination
  - Shadows
  - Complex and hard-to-describe objects
    - Objects from outdoor scenes
    - Non-rigid objects
    - Objects occluding other objects

Introduction

- State of computer vision → The general computer vision problem is unsolved
  - Develop a visual system as good as humans. No progress for 40 years.
- A lot of progress in specific computer vision problems
  - e.g. face recognition used in digital cameras, surveillance, security.
  - e.g. pick and place.

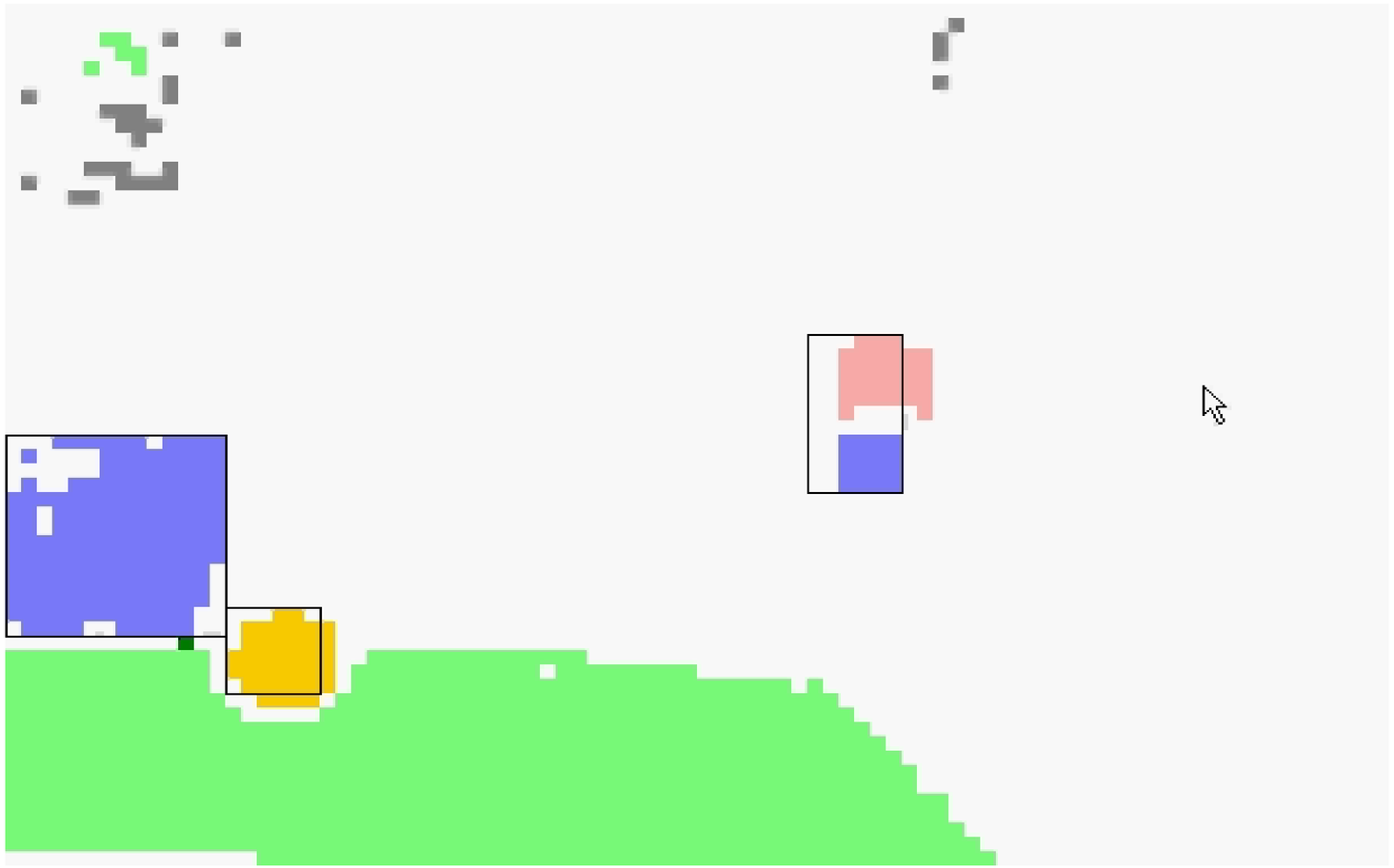
Introduction

Doggie cam



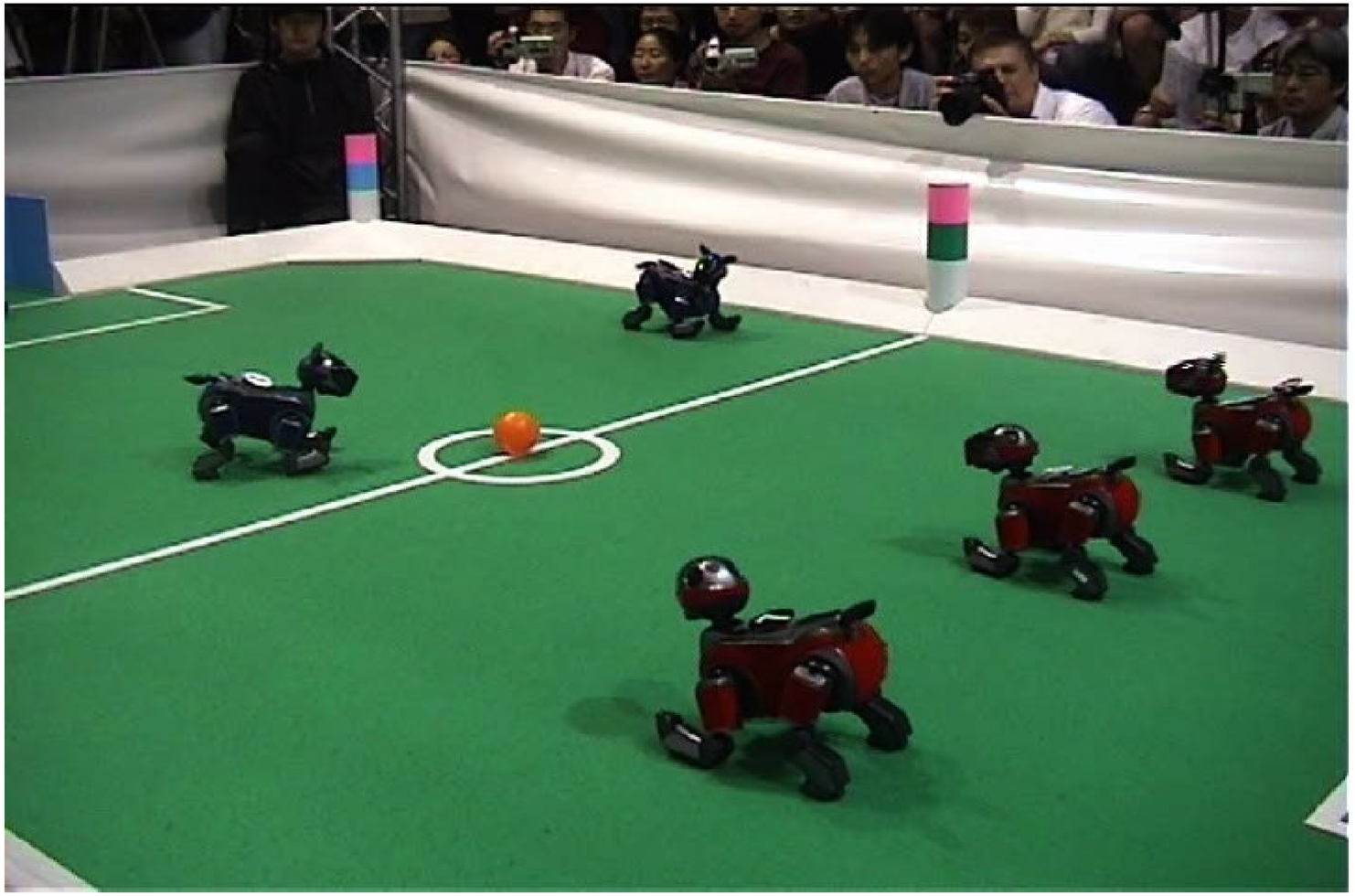
Introduction

Object recognition



Introduction

Computer vision in action

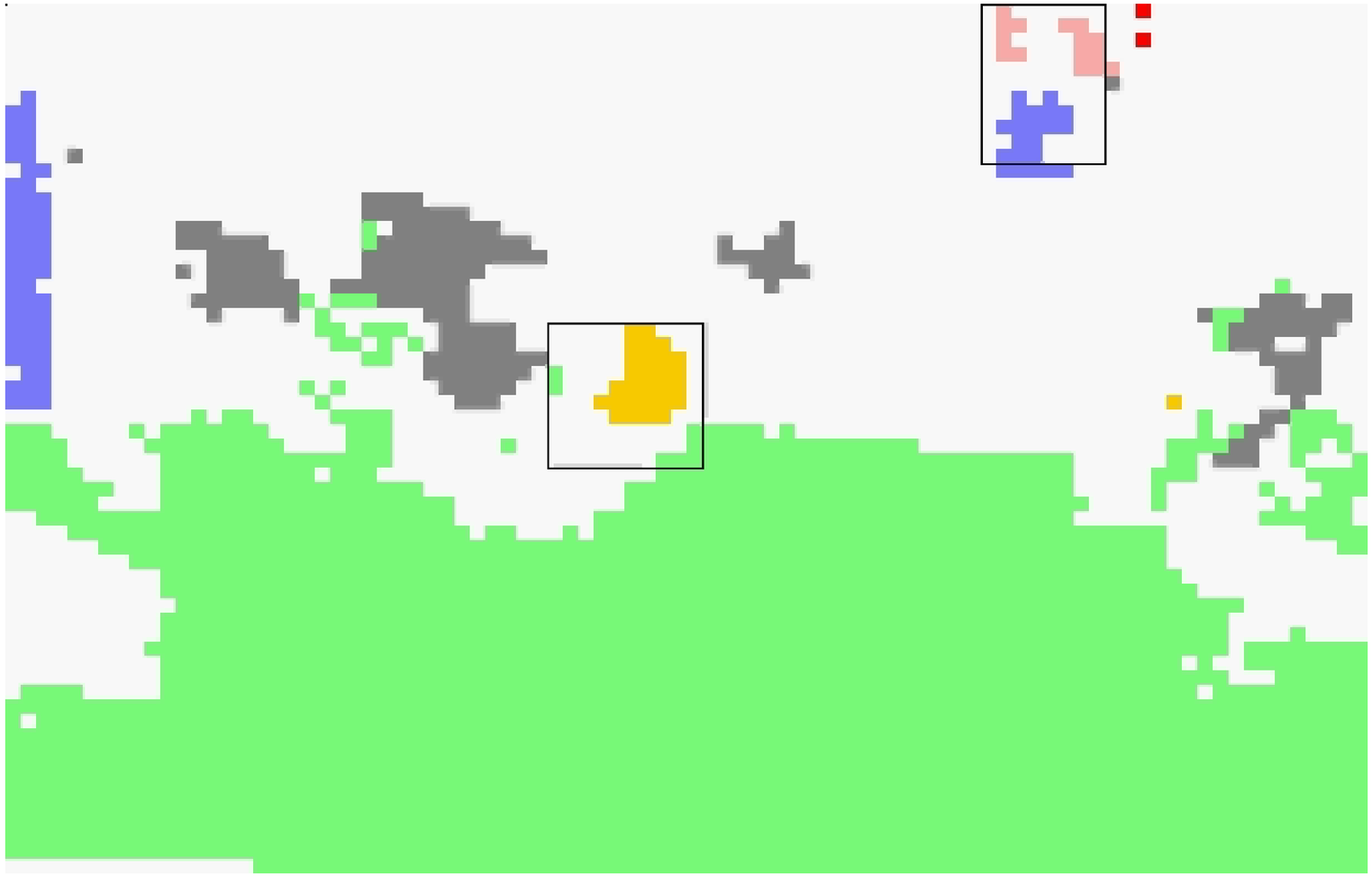


Introduction

Doggie cam in action

Introduction

What the robot sees



Introduction

- Computer vision creates an image of a scene on an array.
- It uses the lens camera to produce a perspective projection of the scene within the camera field of view.
- Perspective projection is a many-to-one transformation.
- It can be noisy due to low ambient light levels.

Introduction

- Perspective projection → Many-to-one transformation.
  - Several different scenes can produce identical images.
  - The image cannot be directly “inverted” to reconstruct the scene.

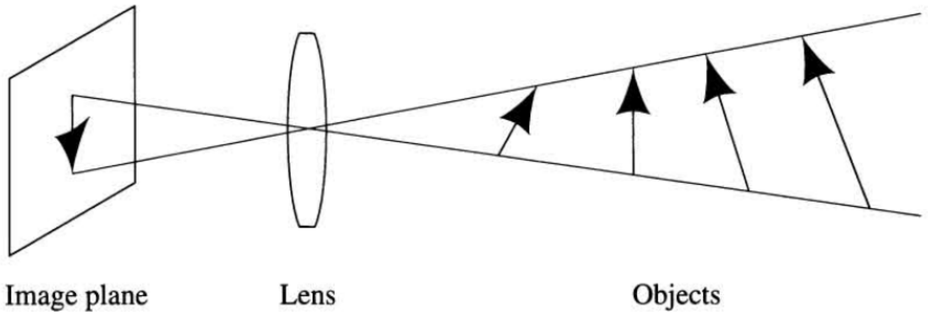


Figure 6.1

The Many-to-One Nature of the Imaging Process

Introduction

- Image is represented as a two-dimensional, time-varying matrix of intensity values I(x,y,t).
  - Colour vision uses three matrices, RGB. Monochromatic only one.
  - In static scenes, time variable is not considered.
- Iconic model or features can be obtained from this matrix.
- Information to be extracted depends on the task, e.g.,
  - For safely navigation: object locations, boundaries, surface property.
  - For object manipulation: locations, sizes, shapes, compositions and textures.
  - Others might include colour, belonging to certain classes.

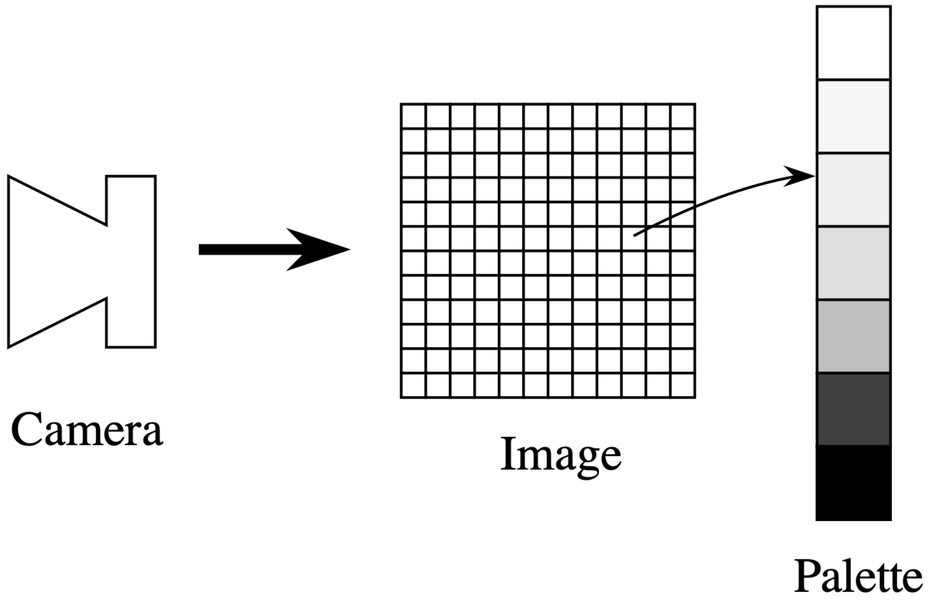
Introduction

- Object features
  - Illumination (incident light)
  - Reflectance (reflected light)
  - Depth (distance from camera)
  - Orientation (angle of normal to surface)
  - Other features: shading, colour, texture

Introduction

Image formation

Light image is created by camera



Each pixel is a number that is an index into a palette of colours or grey scales.

Introduction

- Binary vision
- The original image is “thresholded”, i.e.

$$new[x, y] = (old[x, y] > threshold)$$

- Every pixel brighter than a certain threshold is given a value of 1 otherwise it is zero.
- Easy to process and powerful enough to use in some industrial applications, e.g., picking parts from an assembly line.

Example: Steering an Automobile

- Neural networks can be used to convert the image intensity matrix directly into actions.
  - ALVINN steers an automobile:
  - Input → low-resolution 30x32 image from a mounted camera looking straight ahead.
  - Hidden layer → 5 sigmoid units.
  - Output → 30 units to control the steering angle. Winner-take-all.

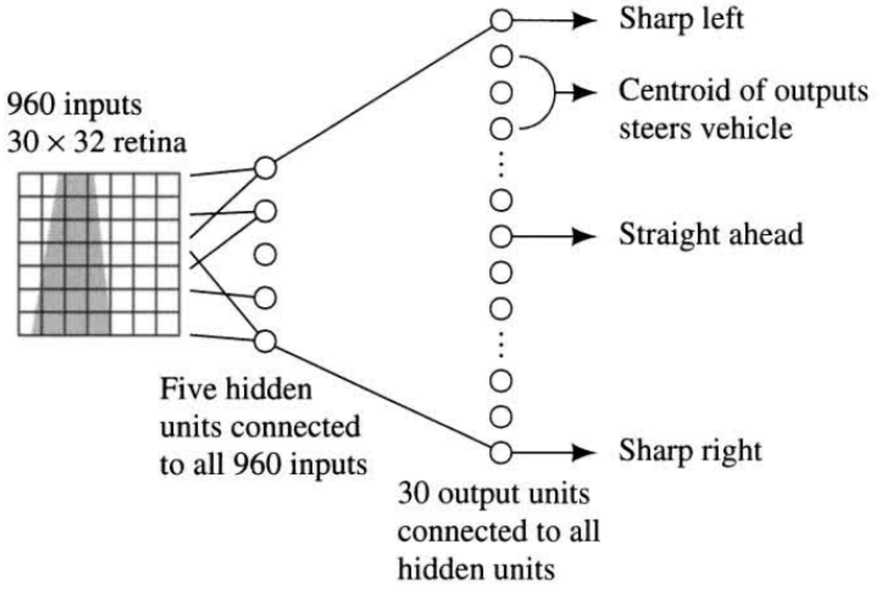


Figure 6.2

The ALVINN Network

Example: Steering an Automobile

- Training 5 minutes of human driving, using actual steering angles as labels.
- Incrementally training with backpropagation.
- Problems:
  - The driver usually drives well.
  - After long, straight distance, the network produces only straight-ahead angles.

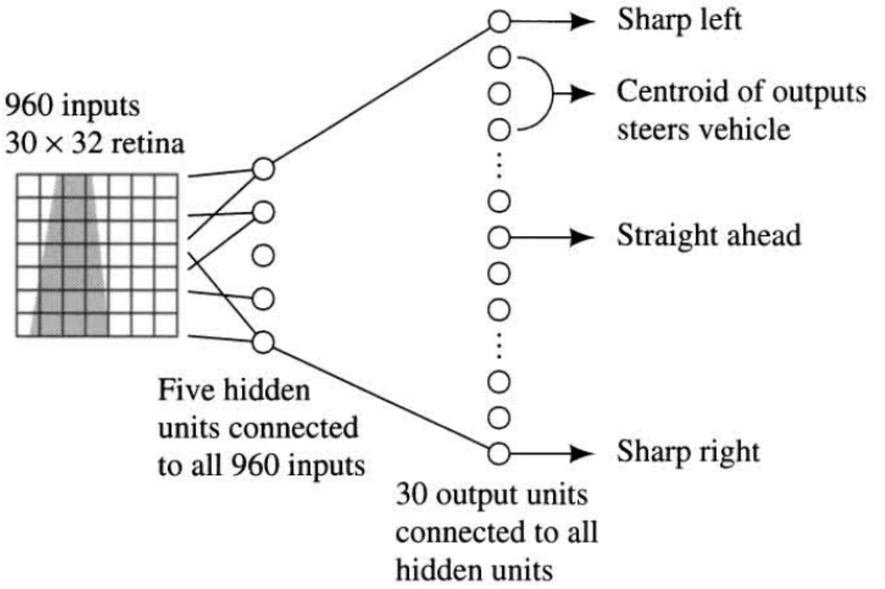


Figure 6.2

The ALVINN Network

Example: Steering an Automobile

- Non-official video:
  - https://youtu.be/oHEH2VDDGss

Robot vision: Two stages

- Man-made environments: doorways, furniture, other agents, humans, walls, floors, etc.
- In exterior environments: animals, plants, man-made structures, automobiles, roads, etc.
- Two techniques: look for **edges** (e.g., intensity changes abruptly) or **regions** (e.g., intensity changes gradually) through **discontinuities.**

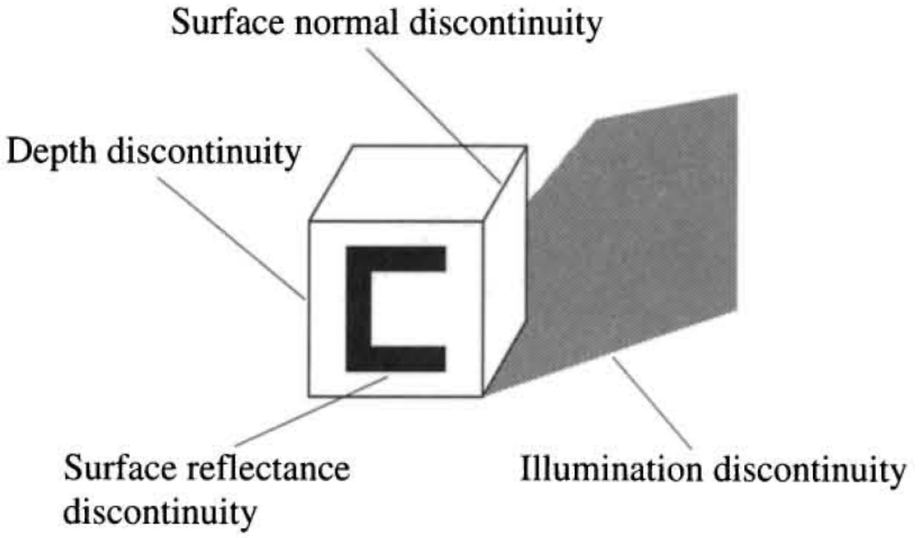
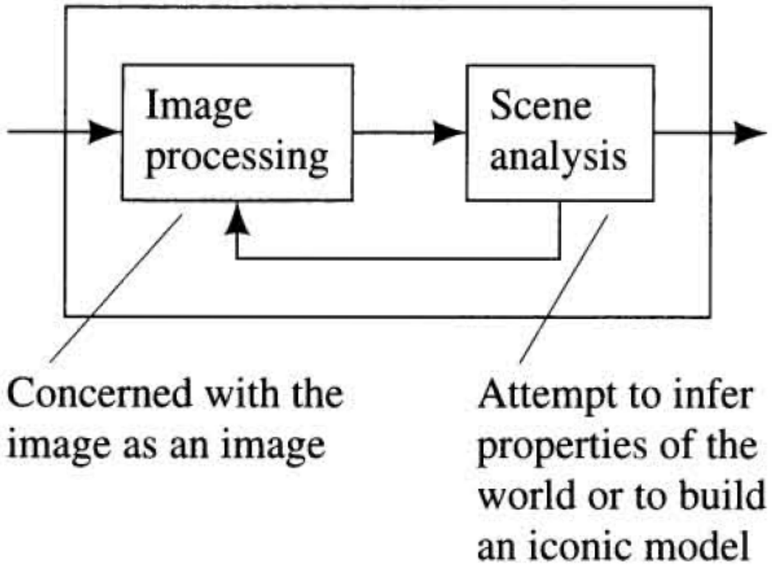


Figure 6.3

Scene Discontinuities

Robot vision: Two stages

- Image processing involves filtering operations to reduce noise, accentuate edges, and find regions.
- Scene analysis creates an iconic model or a feature-based description including only relevant details



Robot vision: Two stages

- From robot view:
  - Three toy blocks (A, B, C).
  - A doorway.
  - A corner of the room.
- Dealing only with disposition of the blocks. Iconic model:
  - ((C B A FLOOR))
  - If C moved → ((C FLOOR) (B A FLOOR)) or ((B A FLOOR) (C FLOOR))

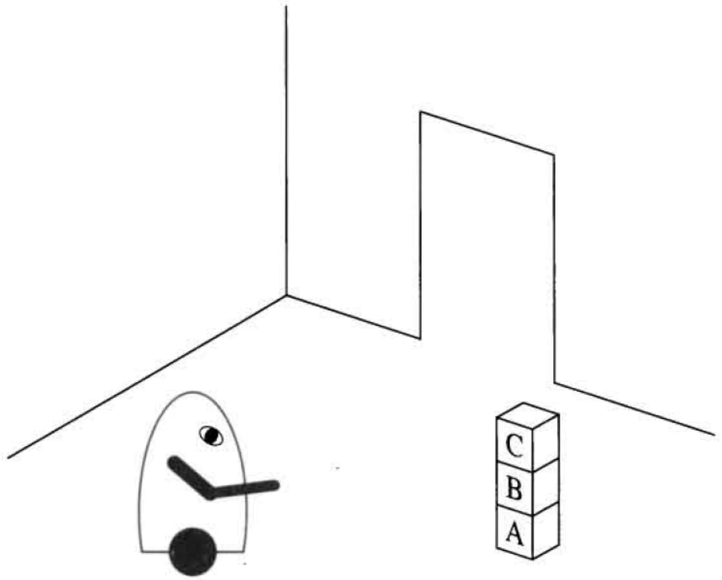


Figure 6.5

A Robot in a Room with Toy Blocks

Lecture Overview

- Introduction
- Image processing
- Scene analysis
- Cognitive vision

Image processing: Averaging

- Image represented as n x m array I(x,y) → image intensity array.
- Cells are called pixels. Each number represent light intensity.
- Real images always contain noise.
- Smoothing tries to remove isolated bright and dark regions.
- Averaging + sliding → convolution.
- Has side-effect of blurring image.

Image processing: Averaging

- It can use a threshold.
- Larger rectangles achieve more smoothing.
- Broad lines are thickened and thin lines eliminated
- In the example, ε = 3, i.e., 0 if sum ≤ 3, 1 otherwise.

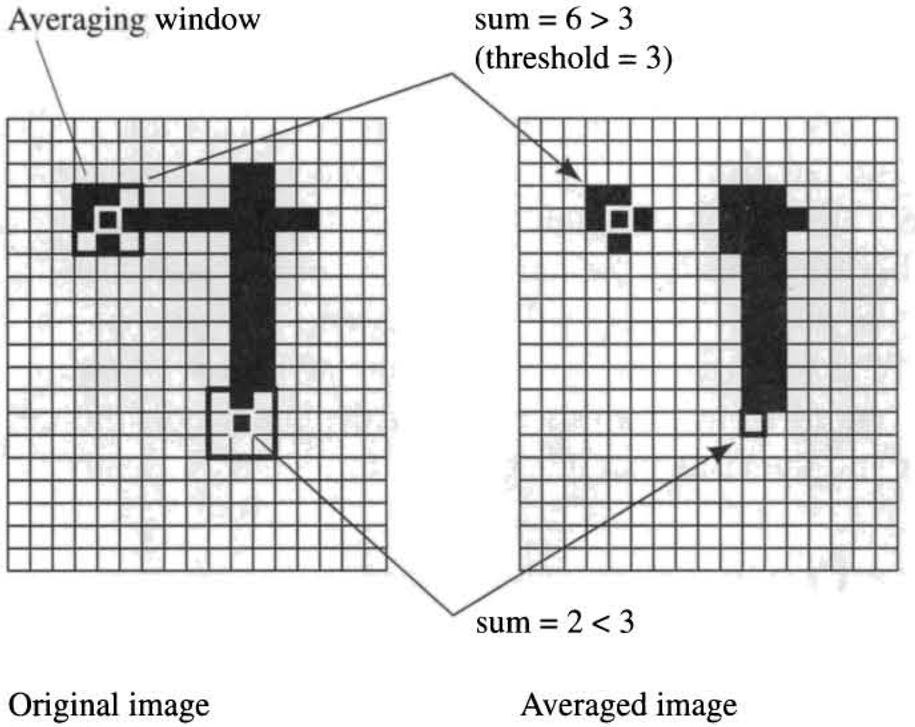
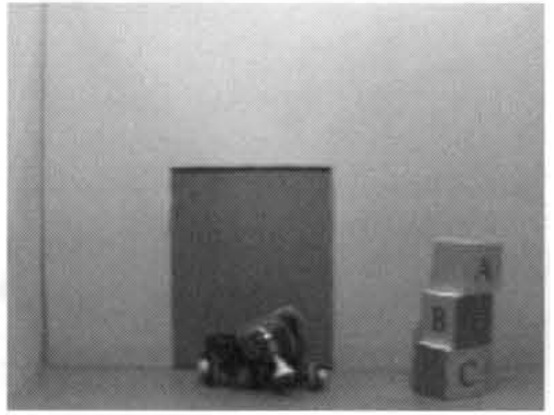
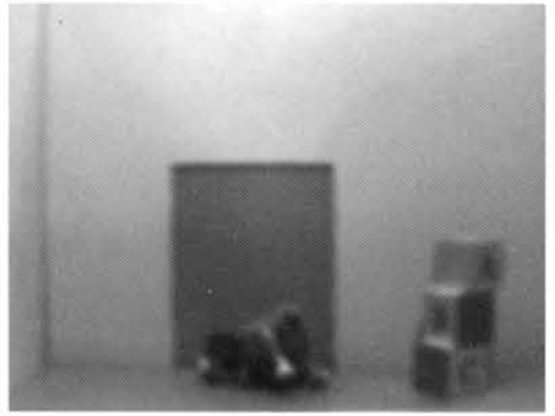


Image processing: Averaging

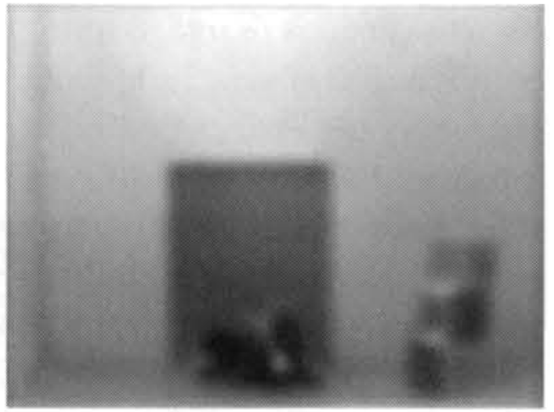
- Image smoothing with a Gaussian filter.
- Images increasingly blurred.



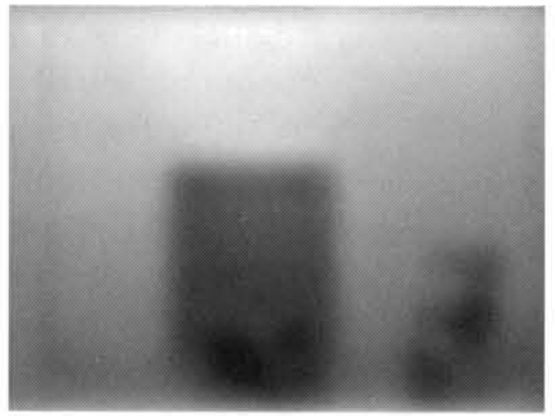
(a) Original image



(b) Width of Gaussian = 2 pixels



(c) Width of Gaussian = 4 pixels



(d) Width of Gaussian = 8 pixels

Image processing: Averaging

- Given a simple 4 x 4 picture matrix:

9	9	9	3
9	9	3	3
9	3	3	3
3	3	3	3

- Smooth this matrix using an averaging technique and a 3 x 3 pixel window.

Image processing: Averaging

- There are four 3 x 3 pixel windows in the matrix.

- Replace middle value in each window by average of all the values in the window.

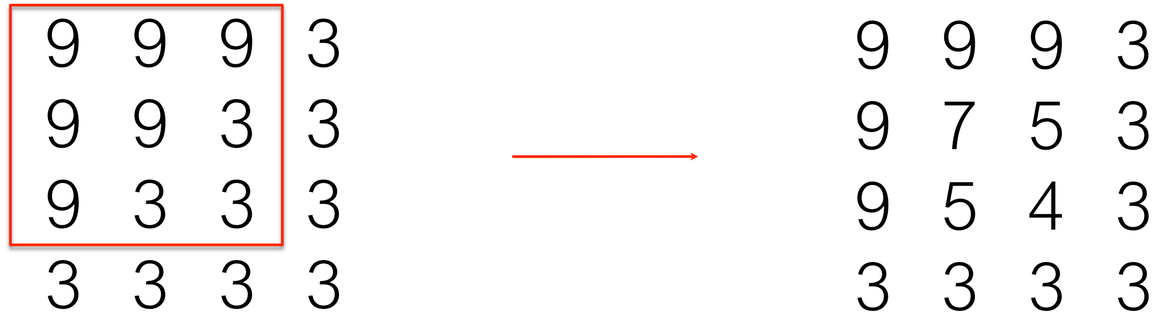


Image processing: Edge enhancement

- Edges are used to build a line drawing.
- Outlines can be compared with object models.
- Edges are parts of the image with markedly different property values (e.g., intensity)

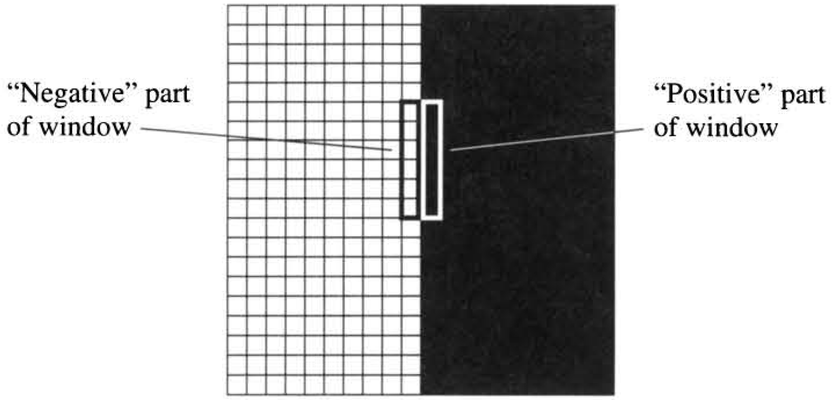


Figure 6.9

Edge Enhancement

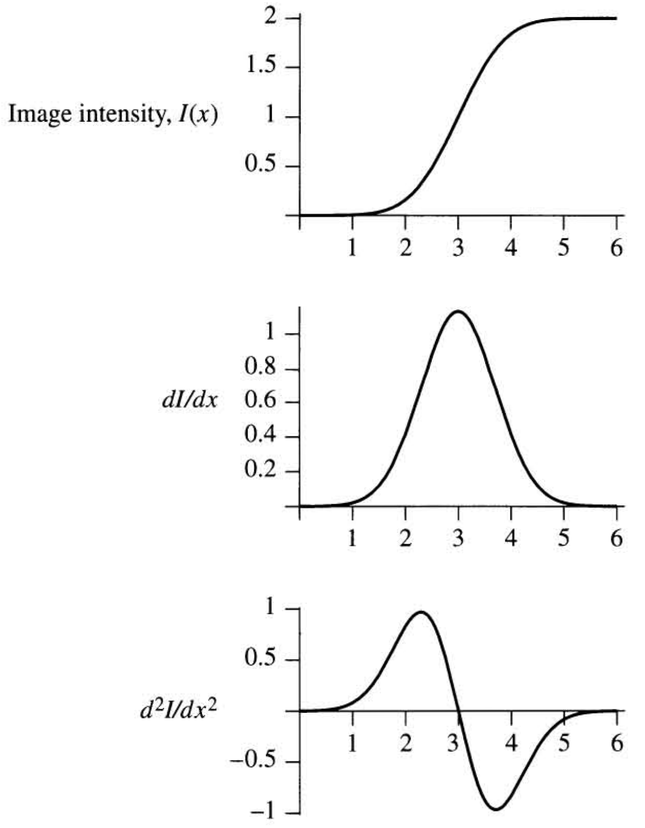
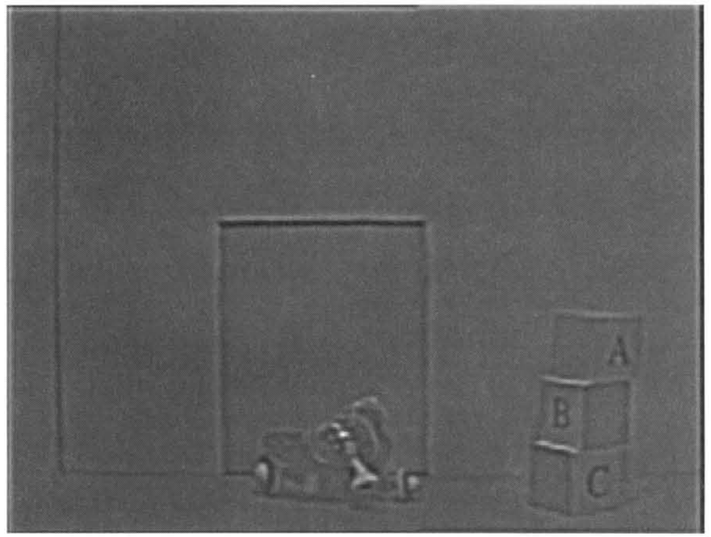


Figure 6.10

Taking Derivatives of Image Intensity

Image processing: Edge enhancement

- Averaging and edge enhancement can be combined.
- For instance, using a Laplacian filter.



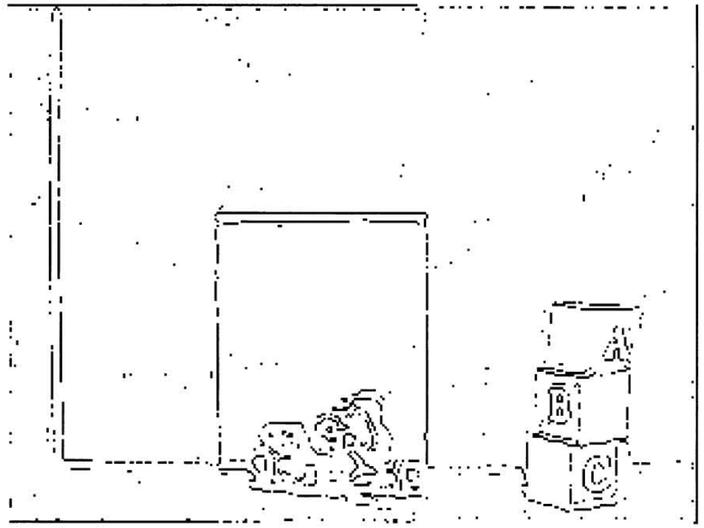
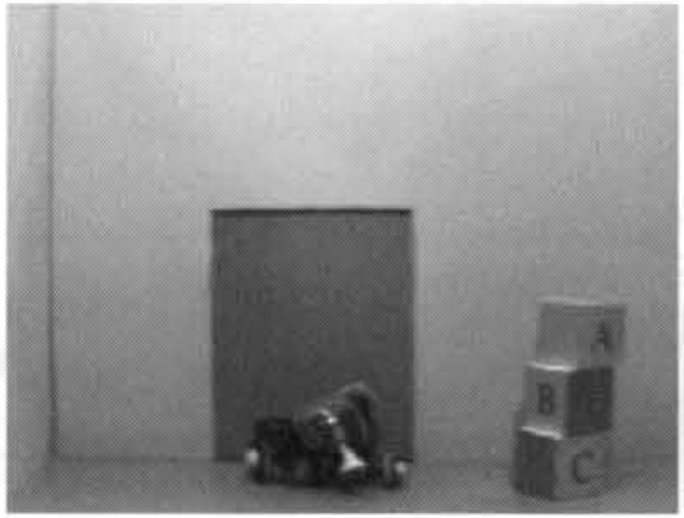


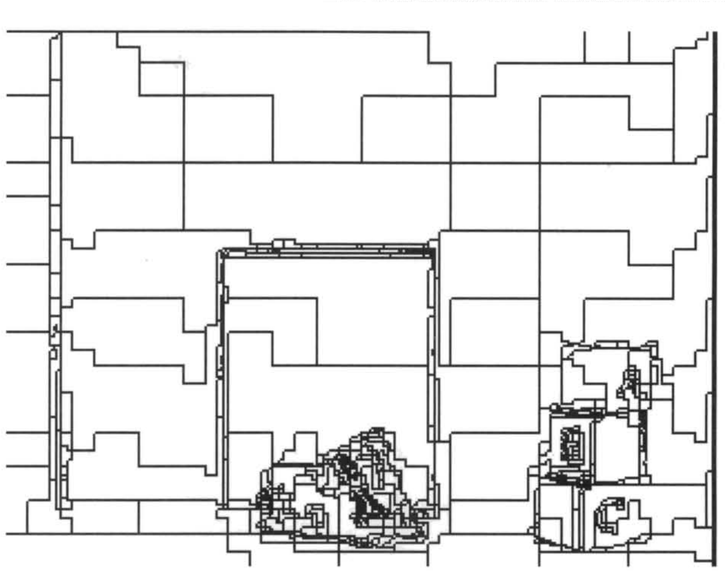
Image processing: Region finding

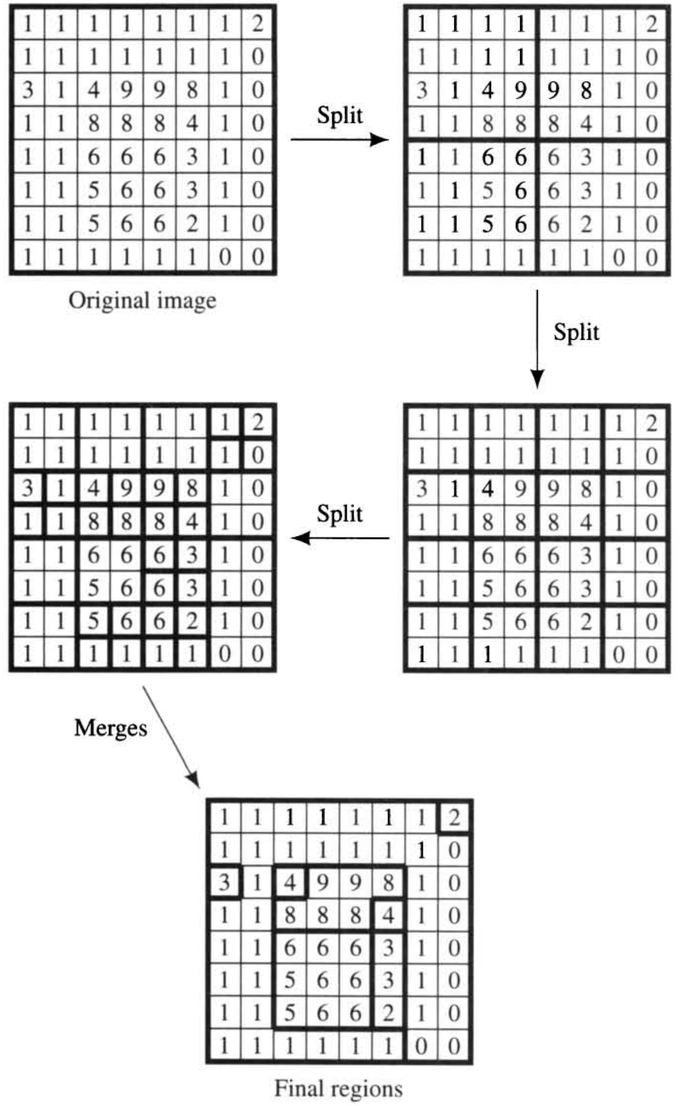
- To find regions in which a property does not change abruptly.
- A region is homogeneous. Intensity difference no more than some ε threshold
- Split-and-merge method. 2^n x 2^n array of pixels.
  - Each no homogeneous region is split in four.
  - Splits continues until no more splits need to be made.
  - Adjacent regions are merged if homogeneous.

Image processing: Region finding

- Splitting and merging candidate regions.
- In this example, intensities may not vary more than 1 unit. Therefore, $\varepsilon <= 1$.







Lecture Overview

- Introduction
- Image processing
- Scene analysis
- Cognitive vision

Scene Analysis

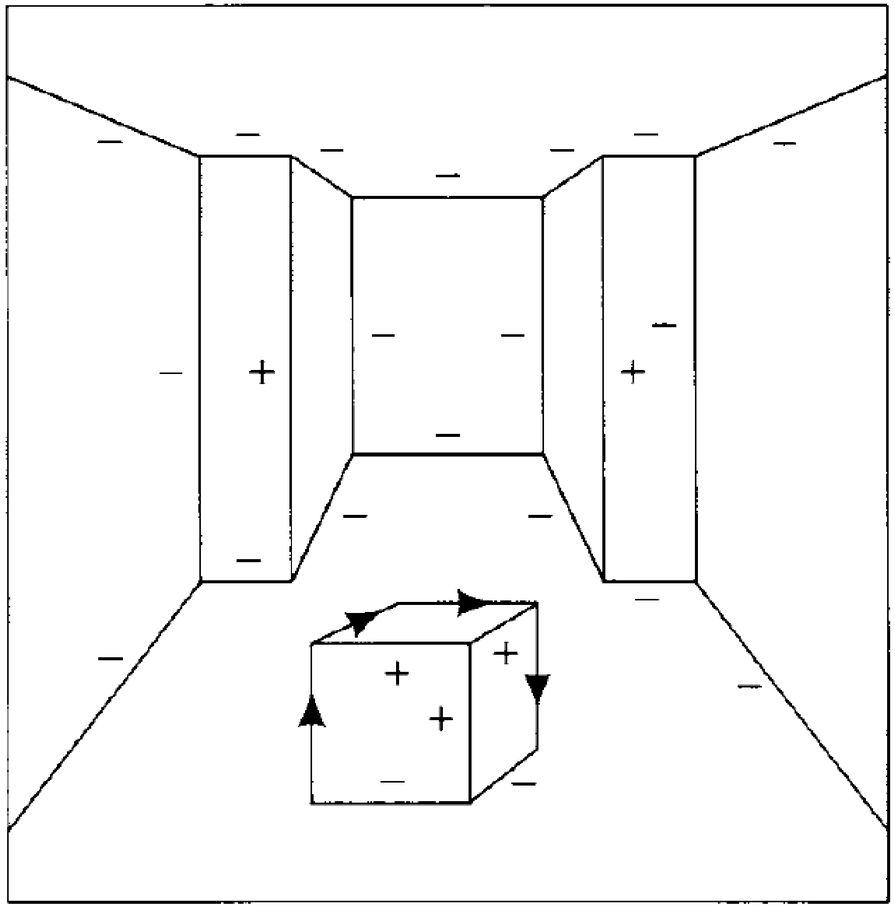
- Extract information about the scene.
- As scene-to-image is many-to-one additional images or information is needed.
- Information can be very general or specific, e.g., camera location, illumination sources, indoors/outdoors, particular objects.
- Iconic model or features:
  - Iconic model builds a model of the scene or part of it.
  - Feature-based analysis is task-oriented.

Interpreting lines and curves in the image

- For scenes with **rectilinear** objects, lines should be postulated. Methods fit segments of straight lines to edges.
- Scenes with **curved** objects fit conic sections (ellipses, parabolas, hyperbolas).
- Interpreting the line drawing associates properties with components of a line drawing.
- For instance, scenes with only planar surfaces have no more than three surfaces intersected in a point.

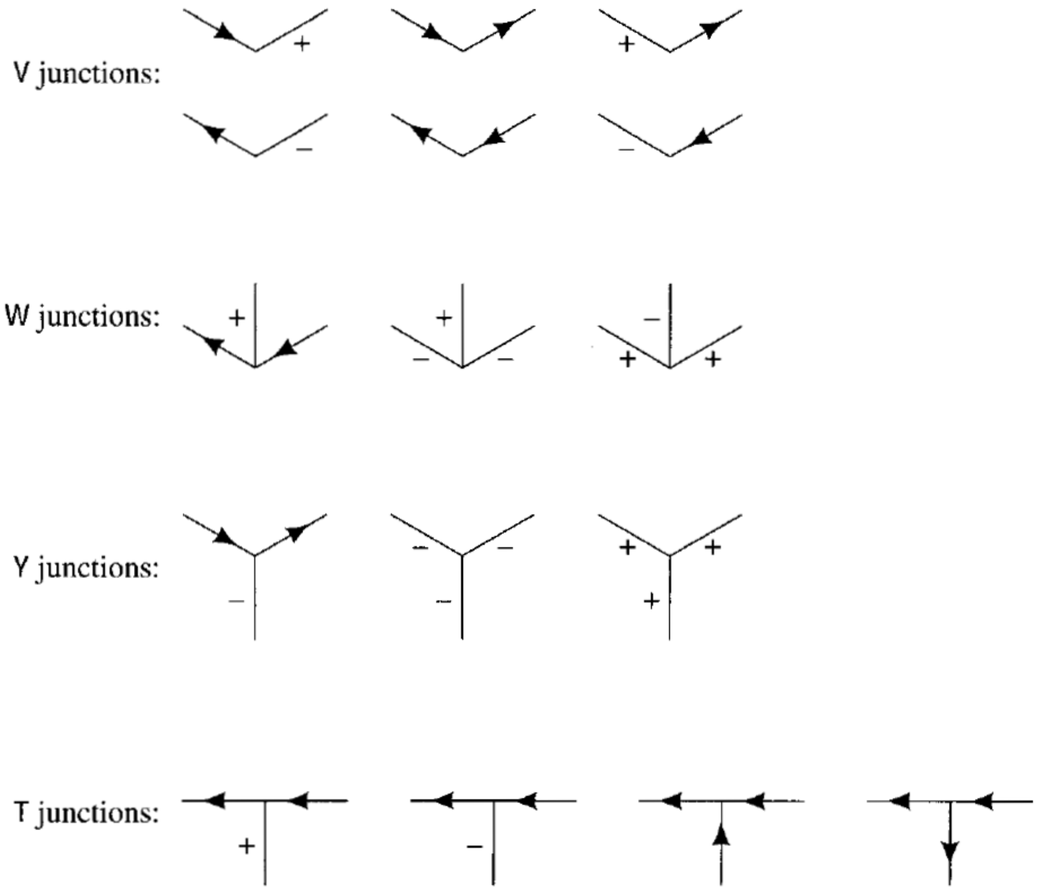
Interpreting lines and curves in the image

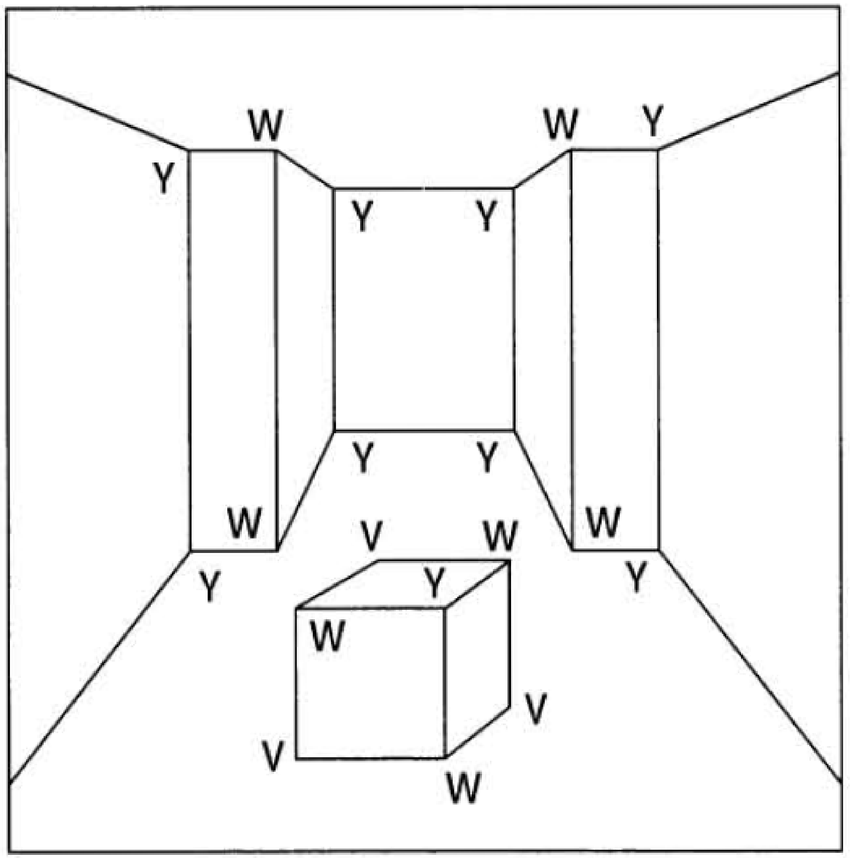
- Scene with bounding walls, floor, ceiling, a cube on the floor.
- Only three possible intersections:
  - Occlude: 2 planes, one occluding (→).
  - Blade: both visible forming a convex edge (+).
  - Fold: both visible forming a concave edge (-).



Interpreting lines and curves in the image

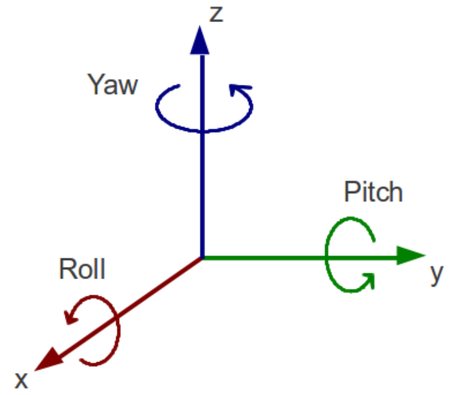
- Labelling types of junctions: V, W, Y, T assigning +, -, →





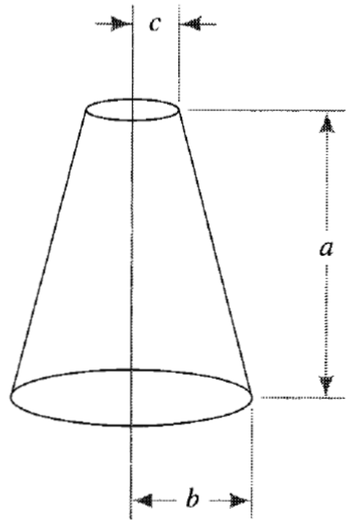
Model-based vision

- Use increasing knowledge about the scene.
- For instance, an industrial scene could use geometric models of components to interpret images – still not semantics though.
- Or if we know a cube is in the scene, a projection can be fitted specifying size, position, and orientation (using Euler angles).



Model-based vision

- Generalized cylinders for model construction.
- Each cylinder uses 9 parameters: a, b, c, 6 location parameters.
- Hierarchical representation.



a, b, c + 6 location parameters

Figure 6.18

A Generalized Cylinder

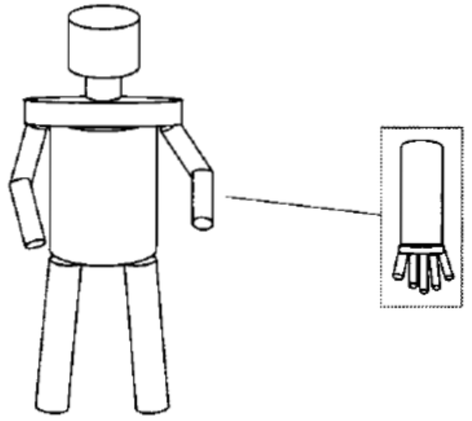


Figure 6.19

A Scene Model Using Generalized Cylinders

Stereo vision

- Under perspective projection large, distant objects might produce the same image as similar but smaller, closer ones.
- Distance estimation from single images is problematic, but sometimes possible.
- e.g., If we know an object is on the floor and the camera height.

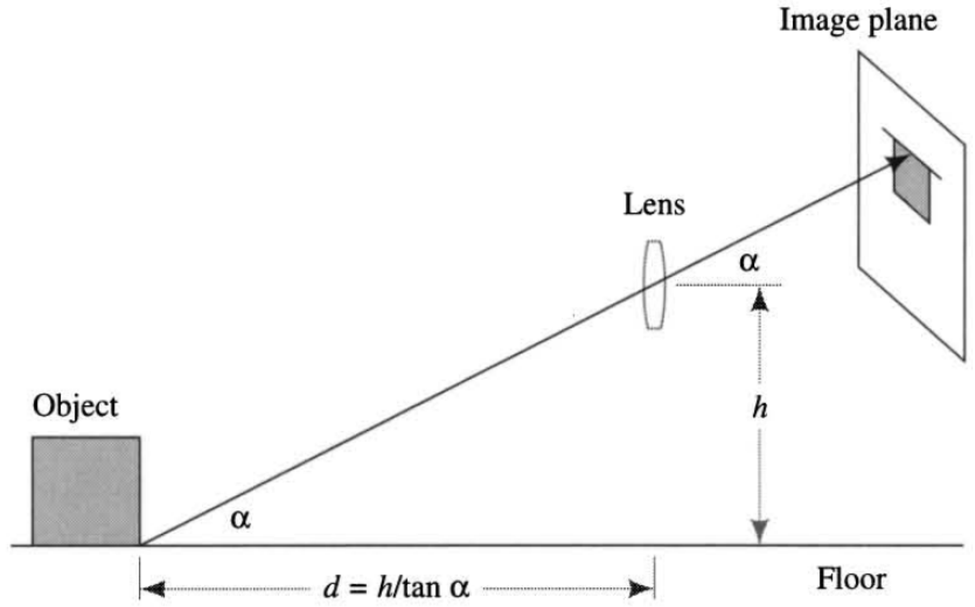


Figure 6.20

Depth Calculation from a Single Image

Stereo vision

- Depth information from stereo vision.
- Two-dimensional setup.
- Two lenses with distance b.
- Correspondence problem for pairs of points.

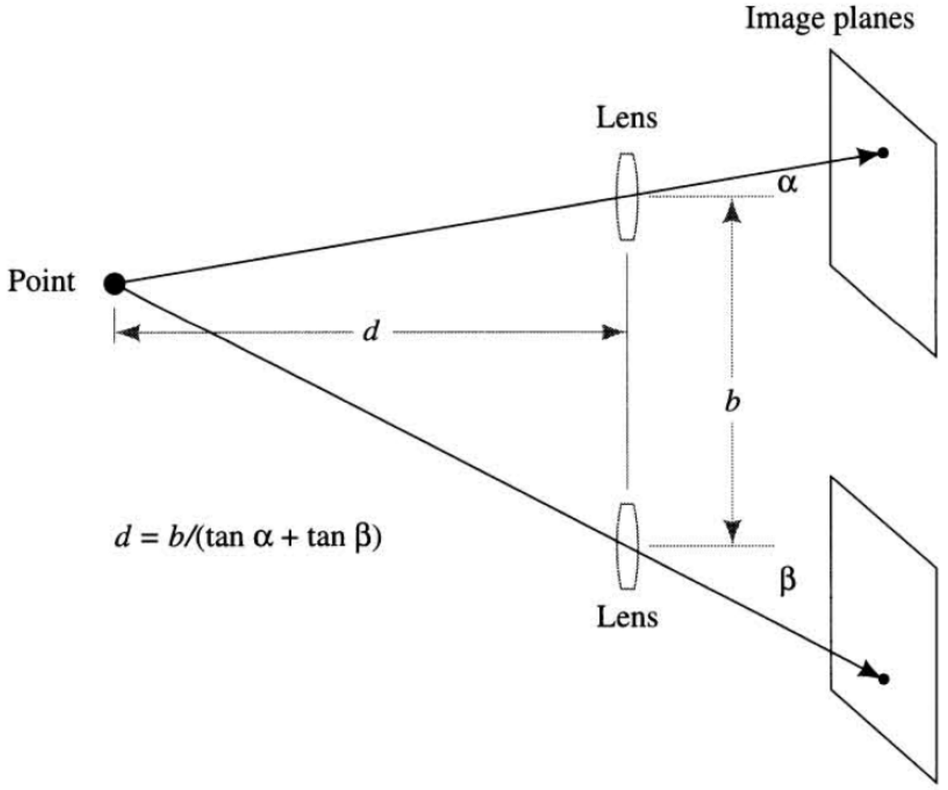


Figure 6.21

Triangulation in Stereo Vision

Lecture Overview

- Introduction
- Image processing
- Scene analysis
- Cognitive vision

Principles of cognitive vision

- Is perception only a recovery process?
  - Computer vision → 3D descriptions of the scene, assigning labels to objects and/or actions.
  - Labels → provided to symbolic reasoning systems.
- Visual perception is seen as a black box delivering labels through recognition, using (mostly static) data.
- From pixels to symbols is difficult, with no causal link between present and past. Therefore, not well-fitted to anticipate the future.

Principles of cognitive vision

- Human behaviour is active!
- Humans (and animals) continuously shift their gaze.
- Humans have intentions and goals linking past with present with the aim of anticipating the future.
- Human actions are goal driven, guided by motor and perceptual expectations.

Principles of cognitive vision

- Is perception only an inference process?
  - Signal analysis is not enough to understand a scene.
  - Additional knowledge through inference – as we look at the world, we think about it.
- Cognitive vision continuously exchanges information between perception and reasoning. A form of predictive vision.
- Actions driven by perceptual expectations – how should I act to see my hand close to the object vs. how should I act to reach the object?

Vision and reasoning interaction

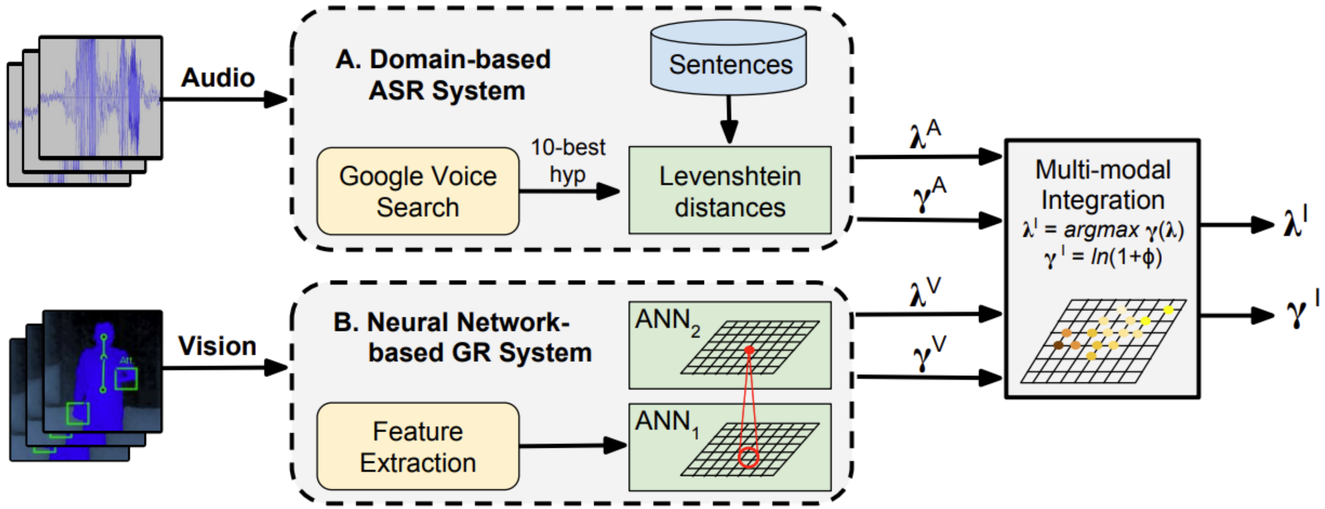
- Cognitive vision extends *processing visual data* beyond the concept of *extracting visual features for real-time control*.
- Reasoning and perception talk about objects, actions, events, and alternative possibilities.
- Loop between prediction (what the system expects perceptually) and exploration (how the system acts to verify if predictions are met).

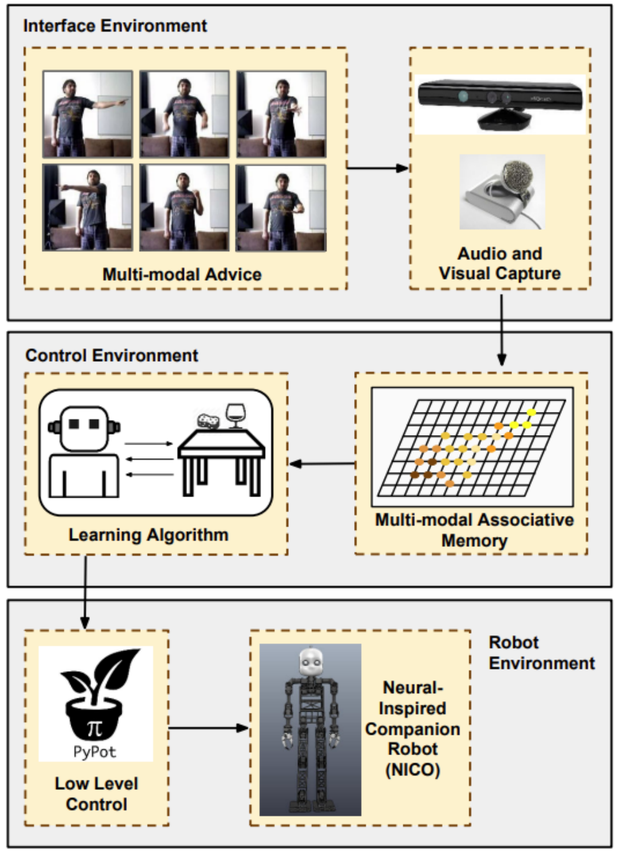
Vision and reasoning interaction

- Five interaction paths for Vision (V) and Reasoning (R)
  - V → R. Traditional perspective for computer vision.
  - R → V. For example “search for the scissors” invokes a visual search.
  - V → R → V. For example “someone is cutting the tomato with the spoon”. Implausible for R so ask V again.
  - R → V → R. For example R needs to know the number of cars.
  - R → VV…V. Imaging and envisioning a situation, action, or event.

Vision and reasoning interaction

- Interactions between V and R can happen at earlier, later, or middle stages.



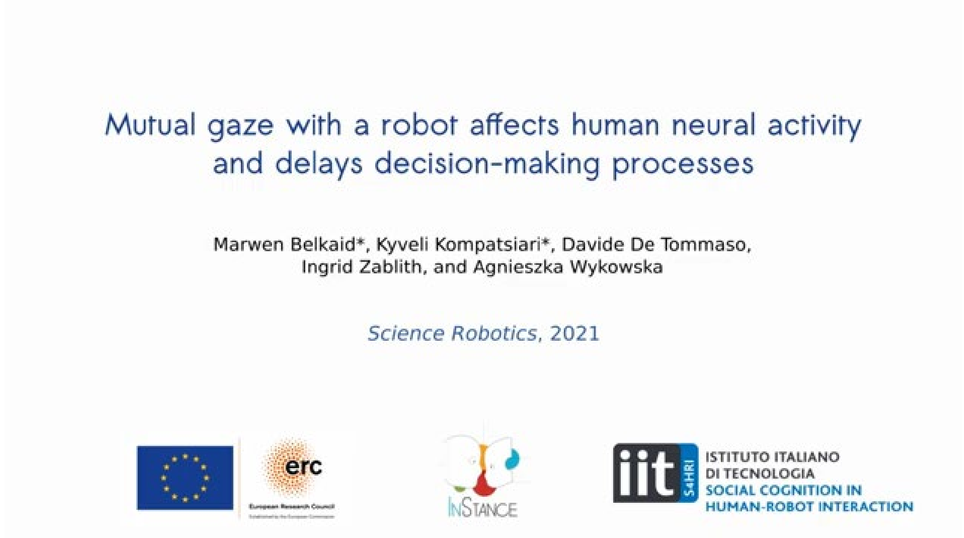


V + R interaction examples

- Cognitive vision to support human-robot interaction.
- iCub’s behavior driven only by the direction of the subject’s gaze making explicit intention to reach for the left or right hand.

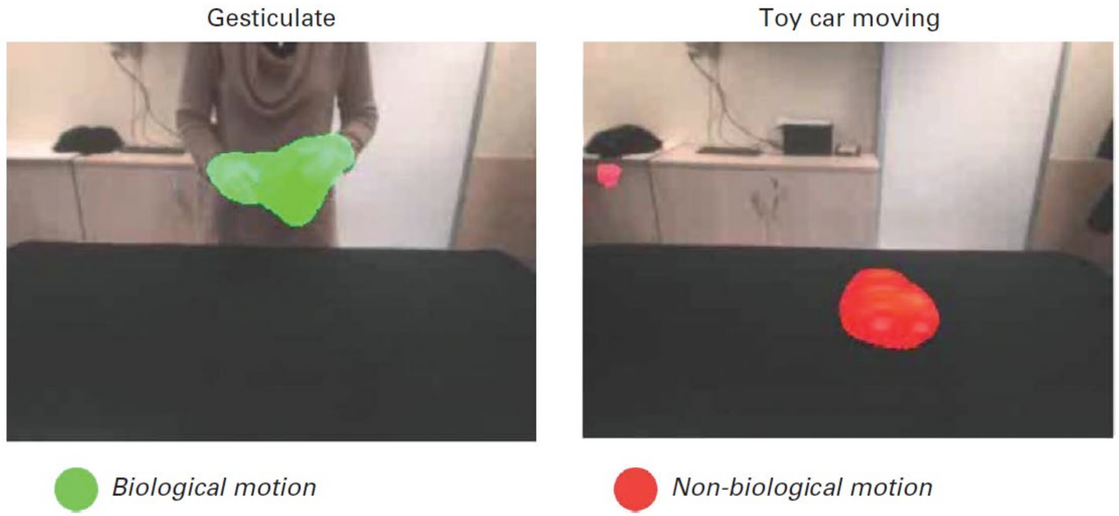
V + R interaction examples

- Cognitive vision to support human-robot interaction.
- iCub's behavior driven only by the direction of the subject's gaze making explicit intention to reach for the left or right hand.



V + R interaction examples

- Cognitive vision for signature of biological motion.
  - → Angular velocity and curvature of the trajectory.
  - Hand during drawing or writing.
  - Knee or ankle during walking.
- Visually measured independently of its shape and color.

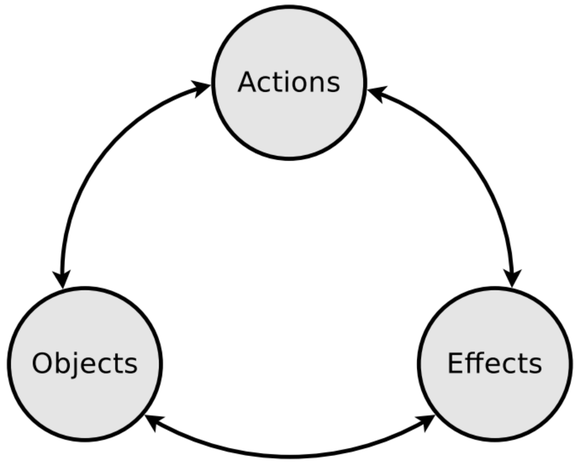


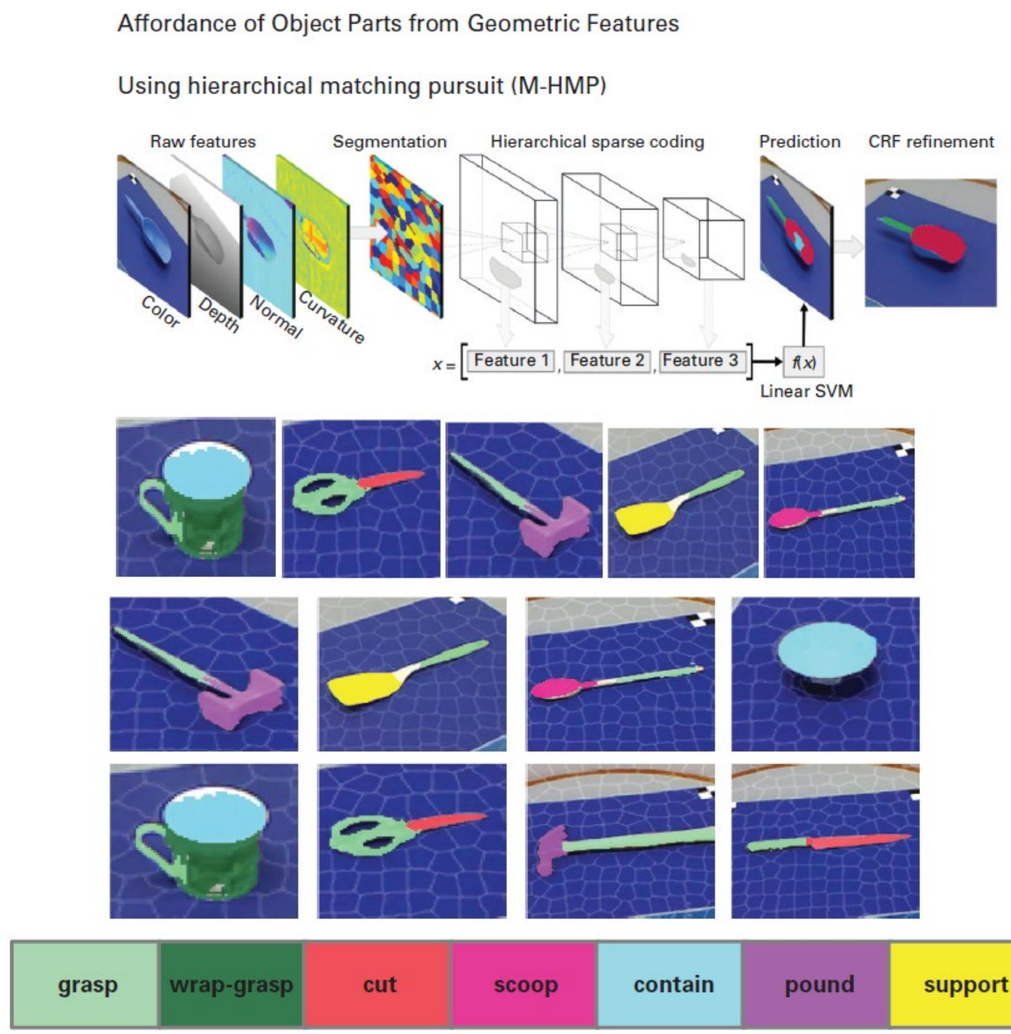
V + R interaction examples

- Cognitive vision involves language as an attention mechanism.
- Synonymy (same meaning) and hypernymy (“is a” relation).
- Objects likely co-occur, e.g., table, cups, spoons.
- A knife put in a drawer is not gone but hidden from the sight. In this case language acts as a part of the reasoning process.

V + R interaction examples

- Cognitive vision for object's affordances
  - Segmentation to infer adjectives.
- Pixel coloured associated to affordance





V + R interaction

- Cognitive vision does not exist in isolation to detect what is where.
  - Direct contrast to how vision is predominantly studied today.
- Unified representation within vision and other sensory modalities through action. V + R → Action.
  - Questions beyond what, where → why, how, who.
  - Also how synthesize visual information to anticipate action effects.

Resources

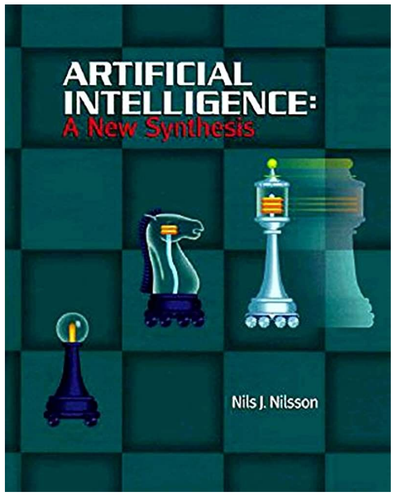
- OpenCV: real-time optimized Computer Vision library.
- https://opencv.org/

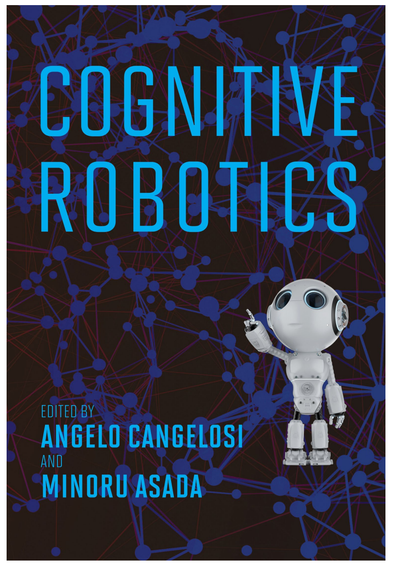
- YOLO: state-of-the-art, real-time object detection.
- https://pjreddie.com/darknet/yolo/

References

- Nilsson, N. J. (1998). *Artificial intelligence: a new synthesis*. Morgan Kaufmann. Chapter 6.

- Aloimonos, Y., & Sandini, G. Principles of Cognitive Vision. In Cangelosi, A., & Asada, M. (Eds.). (2022). *Cognitive robotics*. MIT Press. Chapter 14.





Feedback

- In case you want to provide anonymous feedback on these lectures, please visit:

- https://forms.gle/KBkN744QuffuAZLF8

Muchas gracias!

AI Lecture Feedback

This is a short form to provide early feedback for lectures

franciscocruzhh@gmail.com Switch account

Not shared

* Indicates required question

In case you want a reply, provide your zID. Otherwise your answer is anonymous.

Your answer

how did you participate? *

- In the classroom
- Watch the class from automatic recording

If you have any comments, feedback, or question about the lectures, this is the place. *

Your answer

Submit

Clear form