IIT Jodhpur

Biological Vision and Applications
Module 01-01: About Biological Vision

Hiranmay Ghosh

Vision

- Five sense organs to experience the world:
  - Eyes provide maximum information
- Vision is the process that transforms raw images to information
- This course is about study of principles of biological (human) vision
  - With an ulterior motive of applying them to computer vision system

What does Human Vision System do?

Transforming visual signals to information

Colors
Shapes
Textures

Flower
Rose
Leaves

Image

Description

... This looks trivial !!

What does Human Vision System do? (contd.)

A more complex example

- Determines structural composition of the scene in 3D
- Visual search – where is my cat?

What does Human Vision System do? (contd.)

A still more complex example

- **Identification**
  - Four players
  - Ball, Goalpost
  - Net, gallery, ...
- **Interpretation**
  - Football game
  - Free kick
- **Prediction**
  - Goal score?
- **Action Selection**
  - Cheer?

... Intuitive and instantaneous for humans. Extremely difficult for computers.

What all are involved?

- Signal processing (raw visual signals)
- Cognition capability
    - Mental focus (decide what is important)
    - Knowledge and experience (for interpretation)

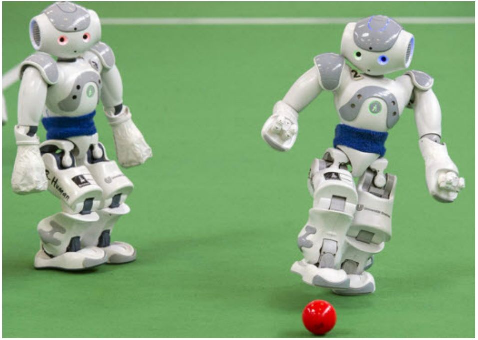
Goal of computer vision

- Simulate human vision system
  - ... Today's CV is far from achieving that
- How to bridge the gap?
  - Study principles of biological vision
  - Create mathematical models for those principles
  - Encode the mathematical models on computers (write programs)
  - Apply those programs for computer vision applications

What are the issues ?

- Too much of data to be handled
  - A HD video camera (1280×780) at 30 fps generates about 90 MB of data per sec.
- Interpreting visual data is ambiguous
  - A red circular object may be a cricket ball or a tomato
  - No two roses are exactly identical!
- Human mind does it intuitively!
  - Data reduction: What to ignore and what to use?
  - Knowledge-based interpretation: Context, knowledge and experience
- Interpretation is subjective

Situated Computer Vision System



- Visual task depends on environment and history
- To be done in real time – followed by action
- Eternal cycle of sensing and action

Layers of Interpretation

Food
Health
Elation

STOP!
Danger

Bharatnatyam
Joy & Freedom

Cognition

Yellow long things
Bananas

Red circle
Traffic signal

Human figure
Outstretched limbs
Blue dress

Perception

Visual signal

Quiz

Quiz 01-01

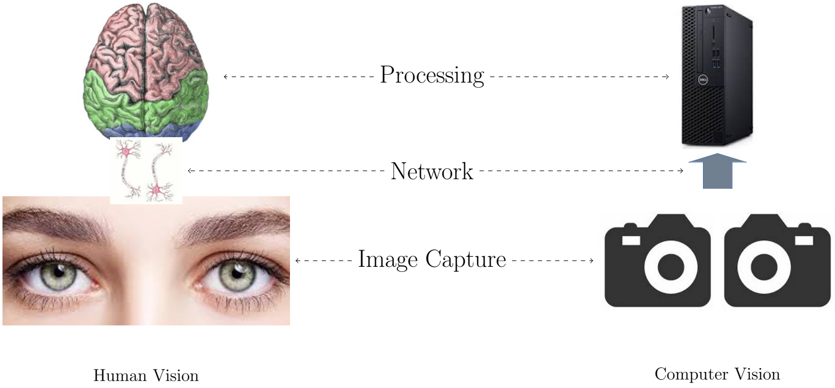
End of Module 01-01

IIT Jodhpur

Biological Vision and Applications
Module 01-02: Stages of Human Vision

Hiranmay Ghosh

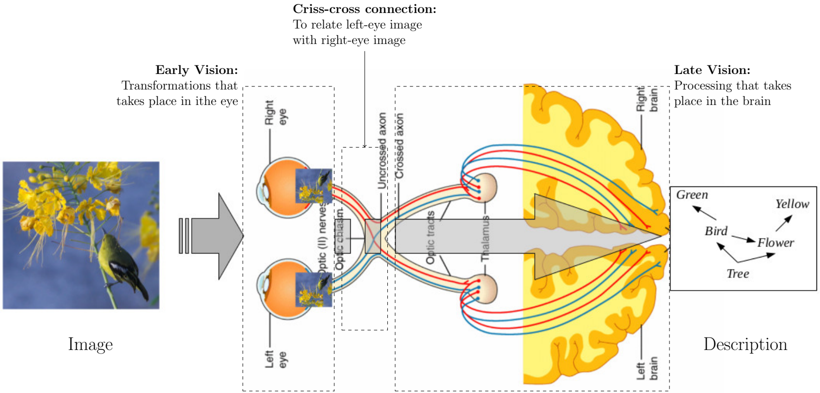
A Simplistic Analogy



… But, there is much more to it

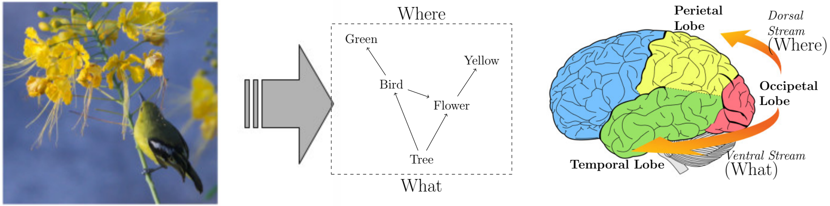
Overview of Human Vision System

Early Vision and Late Vision



Ventral and Dorsal Streams

Answering “What” and “Where”



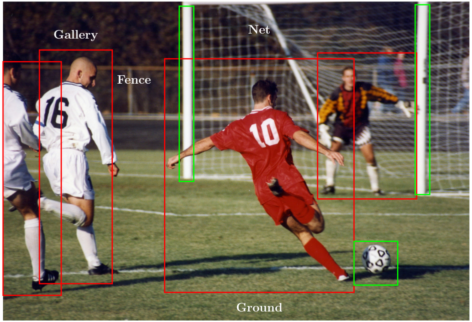
- Ventral Stream is responsible for answering “What”
- Dorsal Stream is responsible for answering “Where”

What is “Vision”?

- **Recognition and localization**
  - Four players, Ball, Goal post
  - Net, viewers’ gallery, fence ...
- **Semantic interpretation**
  - Football game
  - Free kick
- **Prediction**
  - Possible goal score

Perception

Identify objects, locations: What and where



- Foreground and background
- **What ?**
  - Four Players, Ball, Goal Posts
  - Net, Fence, Gallery, Ground ...
- **Where ?**
  - Relative positions
  - Geometric organization

What is perception

- Interpretation of the sensory data - signal processing
- Some assertions made about the environment
- From signals to semantic representation
  - Results in data reduction
- Different viewers can make different assertions about the same scene
  - Depending on viewpoint, signal noise, sensory capabilities, etc.
- May be correct, partially correct, or incorrect
  - There can be “illusions”

Attention

Decide what is important

- We never look at the whole picture
  - We look at selective places for understanding the scene
- Lots of information gets filtered out before entering cognitive system
- Visual semantics is conveyed by very small regions of a picture

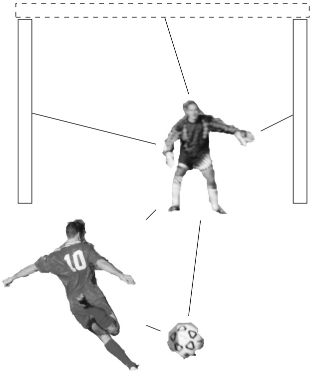
What is attention

- Selective filtering of the sensory data
  - Deciding what is important
- Results in data reduction
- Depends on context, user intention, task at hand
- Can result in change blindness
  - Adversarial attempt to "divert" attention

Edpuzzle assignment

Cognition

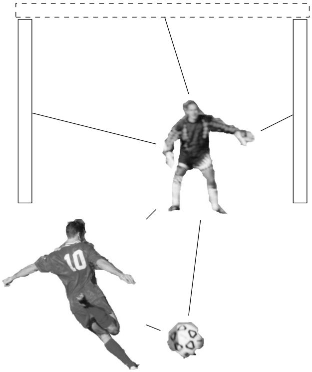
What do you "infer" from this picture ?



- Focus on a few things and interactions
- Draw inferences:
  - It is a football game
  - It is a free-kick
  - The goal-keeper is ready to defend
- How do you make these inferences?
  - Knowledge of football and other games
  - Experience of this game, reputation of the players ...

Prediction

What do you “predict” ?



- What will be the likely trajectory of the ball?
- What will be the goal-keeper’s reaction?
- What is the probability of a goal score?
- How do you make these inferences?
  - Knowledge of football and other games
  - Experience of this game, reputation of the players ...

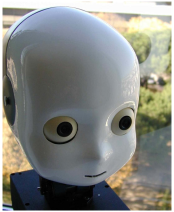
What is cognition

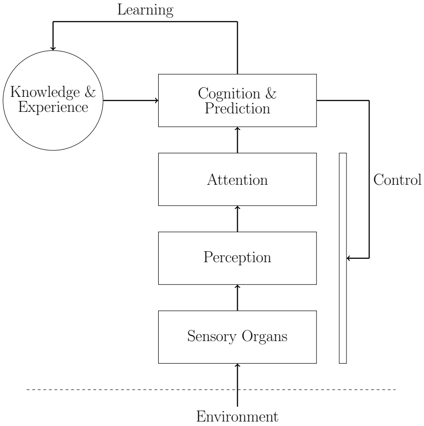
- Experiential interpretation of the filtered percept
- Cognition is subjective, depending on
  - Intentional state of the observer
  - Background knowledge, context, experience, etc.
- Can fill-in the missing percept / correct erroneous perceptions
- Cognition includes prediction (past and future)

In summary ...

- **Perception:**
  - Acquisition of new information about the environment through the sensors
  - Sensory signals to symbolic representation
- **Attention:**
  - Selective filtering of percept (decide what to filter)
  - Depends on user task, intention, context, etc.
- **Cognition:**
  - Experiential interpretation of filtered sensory data
  - Inferencing about the (past / present / future) state of the world

Simplified process model for Vision System





Quiz

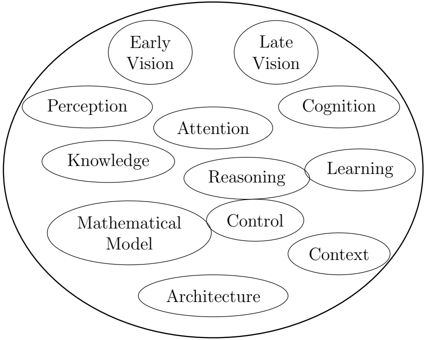
Quiz 01-02

End of Module 01-02

Biological Vision and Applications
Module 01-03: About the Course

Hiranmay Ghosh

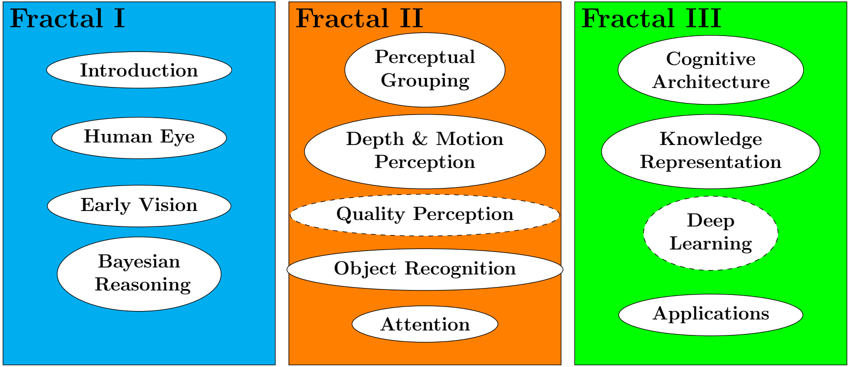
We have introduced the important concepts for the course



Biological Vision

... We shall elaborate them and relate them during this course

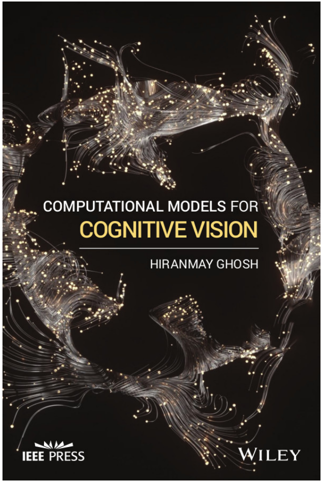
The course is divided into three fractals



Focus and prerequisites

- More focus on Computer Science/AI topics
  - AI, ML, Mathematical modeling, Computer Vision ...
- Results of psychological experiments will be discussed.
- Prerequisites
  - **Mathematical/statistical skills**
  - Computer Vision: Desirable, but not necessary
  - AI / ML: Will be introduced as necessary
  - Psychology / Neurology: No

Study Material



- Textbook:
  - Hiranmay Ghosh.
    Computational Models for Cognitive Vision.
    Wiley-IEEE Press, 2020.
  - Access link to be put up in classroom
    - No download
- Research papers will be announced in the class

Evaluation

Class assignments

- **Continuous Evaluation [60]**
  - Simple quiz at the end of every class (well almost!) [20]
    - Immediate deadline (No second chance)
  - Programming / non-programming assignments [40]
    - Use C/C++/Java/Python – Colab recommended (No exotic language)
    - EdPuzzle assignments
- **Examinations [40]**
  - Major [20]?
  - Fractal end examinations [20]?

Plagiarism Policy

- Will be severely dealt with
- First offense: Zero marks for the complete assignment
- Second offense: Report to institute for appropriate action

Quiz

Quiz 01-03

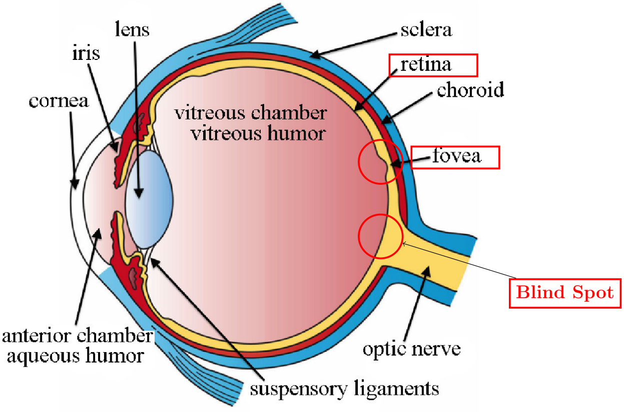
End of Module 01-03

IIT Jodhpur

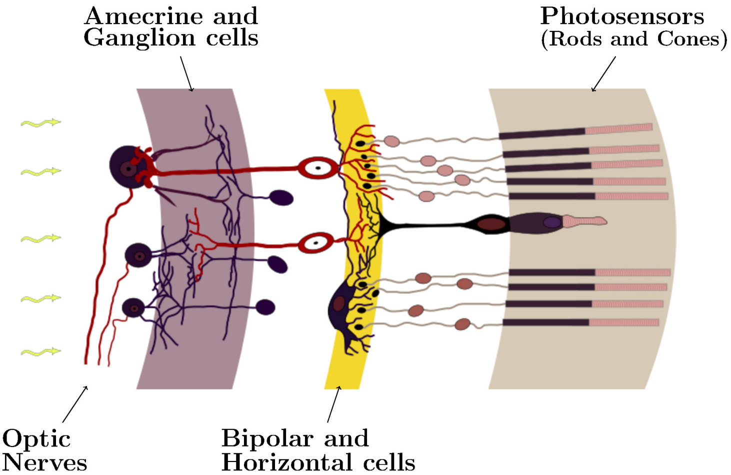
Biological Vision and Applications
Module 02-01: Structure of the eye

Hiranmay Ghosh

Structure of the eye



The Retina

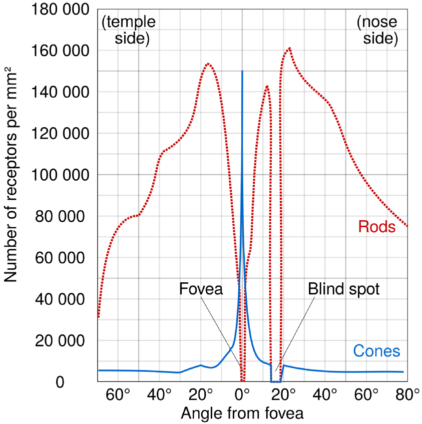

Amecrine and
Ganglion cells
Photosensors
(Rods and Cones)
Optic
Nerves
Bipolar and
Horizontal cells

The Photosensors

Rods and Cones

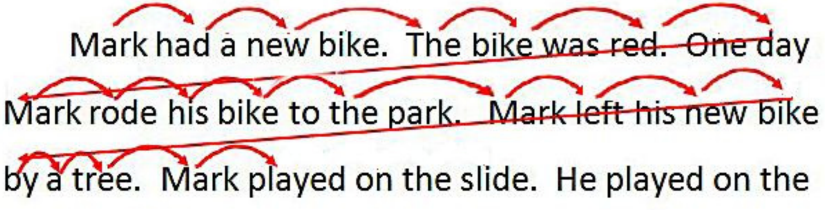
- Rods
  - More sensitive to light, but insensitive to Color
  - Responsible for night (low light) vision
  - About 120 million rods in each eye
- Cones
  - Enables color vision
  - Three types of cones with different sensitivity to wavelengths
  - About 6 million cones in each eye

The Photosensors are not uniformly distributed over the retina



- Cones cover about 15% of visual field,
  - This area is called the <u>fovea</u>
  - Maximum acuity in this area
    - Best within 1.5 – 2 degrees
- Rods spread over 60 – 80% of visual field
  - This area is called the <u>peripheral</u> area
  - Resolution decreases linearly with the distance from the center of the eye

Attention, Fixation and Saccade



- Attention:
  - Orient the eye so that the object of interest is seen through the foveal area
- Fixation:
  - Intermittent stopping of foveal position on a single location when eye acquires information
- Saccade:
  - Eye movement between two successive fixations

Experiment

Focus your gaze on the red dot and try to read the text

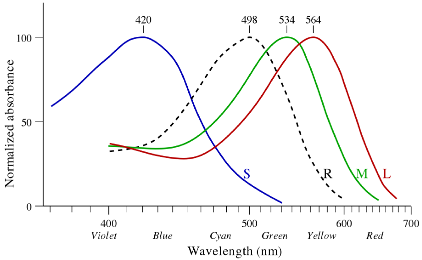
The uneven distribution of the cones on the retina results in high acuity image to be formed only in the foveal region, which covers about $1.5 - 2^{\circ}$ of the visual field. Best acuity occurs at the *fovea centralis* that is about $\frac{1}{10}$th of the fovea. When a person looks at an object, it is brought to the center of the visual field with eyeball movement. The process that controls the eyeball movement is known as *visual attention*, which we shall discuss in chapter 5. The image formed in rest of the visual field is with low acuity and contributes to *peripheral vision*.

Experiment

Most of you should be able to read within the dotted circle

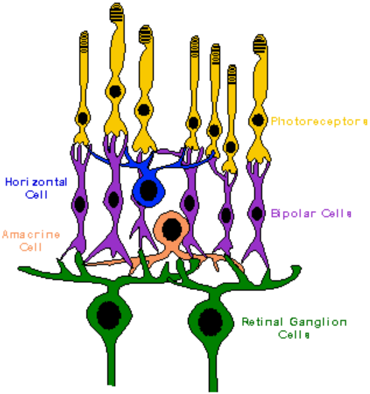
The uneven distribution of the cones on the retina results in high acuity image to be formed only in the foveal region, which covers about $1.5 - 2^\circ$ of the visual field. Best acuity occurs at the *fovea centralis* that is about $\frac{1}{10}$th of the fovea. When a person looks at an object, it is brought to the center of the visual field with eyeball movement. The process that controls the eyeball movement is known as *visual attention*, which we shall discuss in chapter 5. The image formed in rest of the visual field is with low acuity and contributes to *peripheral vision*.

The cones



- Cone-S, Cone-M, Cone-L:
  - Maximum response to short, medium and long wavelengths respectively
  - Response levels of the cones determine color perception
- Rod:
  - Maximum response at medium wavelength

Other cells on the retina



- Horizontal Cells
  - Connects photosensors to neighbors of same kind
- Bipolar Cells
  - Connects photosensors to ganglions
- Ganglion cells
  - Connects to brain via optic nerves
- Amecrine cells
  - Interconnects the ganglions

We shall look at their functions in the following modules

Quiz

Quiz 02-01

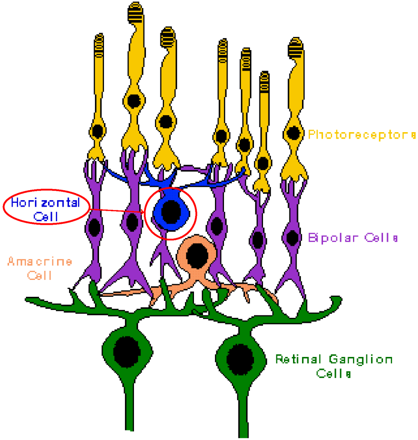
End of Module 02-01

IIT Jodhpur

Biological Vision and Applications
Module 02-02: Lateral Inhibition

Hiranmay Ghosh

The Horizontal cells and Lateral Inhibition

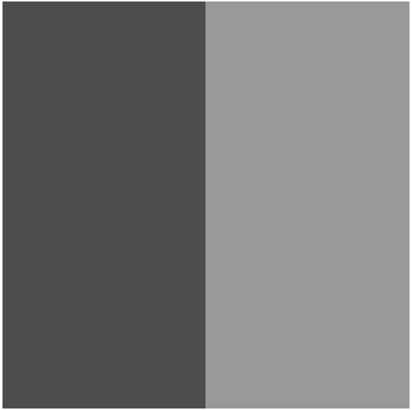


- Connects nearby photoreceptors of the same kind
  - Rods or Cones of same type
- Carries a Lateral Inhibition signal
  - From a photoreceptor to it's neighbors
  - $\propto$ response level of the photoreceptor

Machband Effect

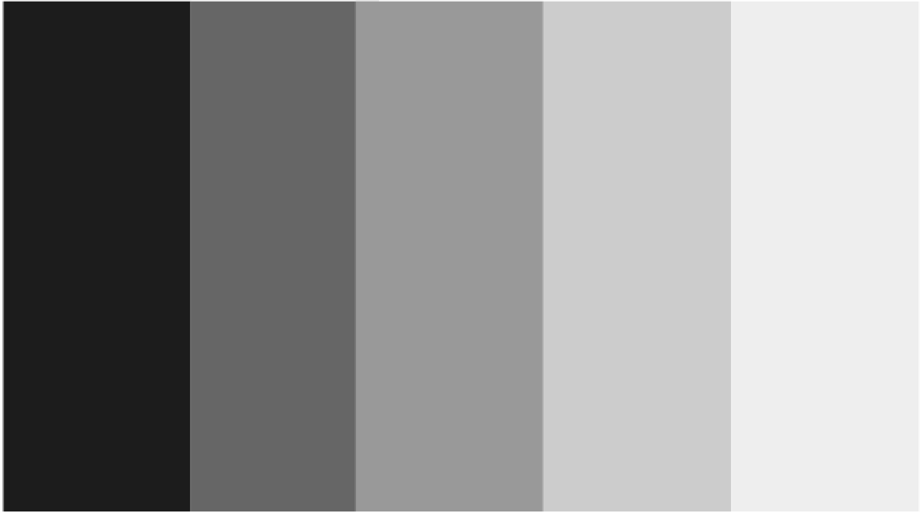
Machband Effect

… contd



Machband Effect

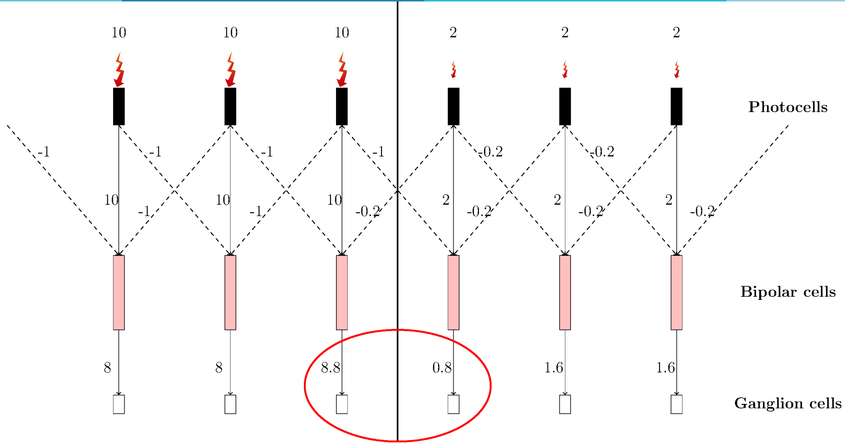
... contd



Explanation

- The bars are of uniform colors
- Color next to a darker color appears lighter, and vice-versa
- Adds contrast to vision
- Happens because of the LI signal from the adjacent photoreceptor cells

Computational Model



- Assumes $\frac{1}{10}$th of response to be transmitted as inhibition signal to neighbors

Quiz

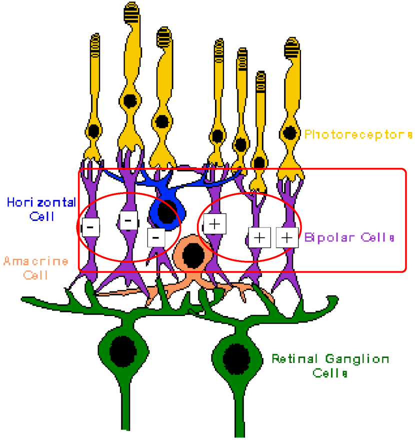
Quiz 02-02

End of Module 02-02

Biological Vision and Applications
Module 02-03: Edge Perception

Hiranmay Ghosh

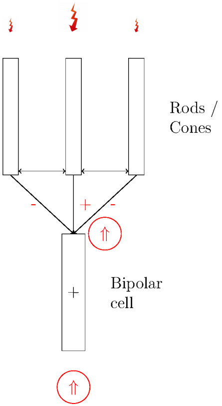
The Bipolar cells



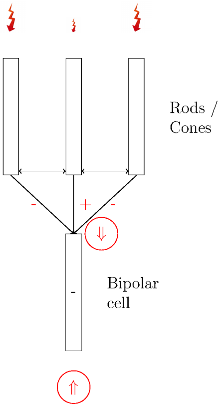
- Photoreceptors are connected to Ganglions through bipolar cells
- Some of them “invert” the signal

The Bipolar cells

On-center and off-center configurations

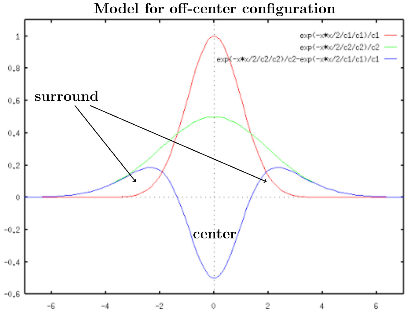


On-center configuration



Off-center configuration

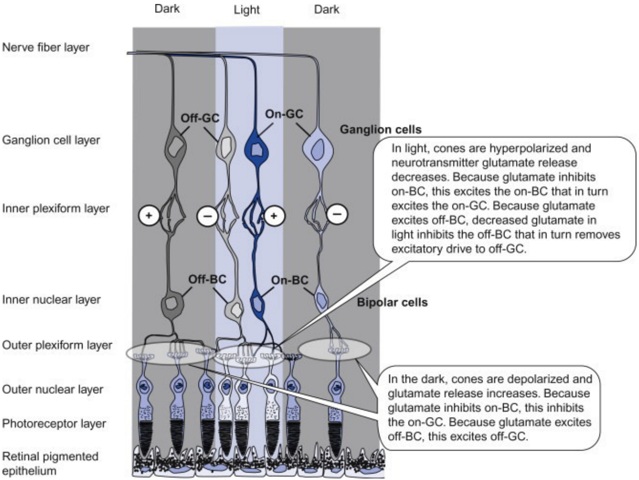
Mathematical model
Difference of Gaussian (DoG)



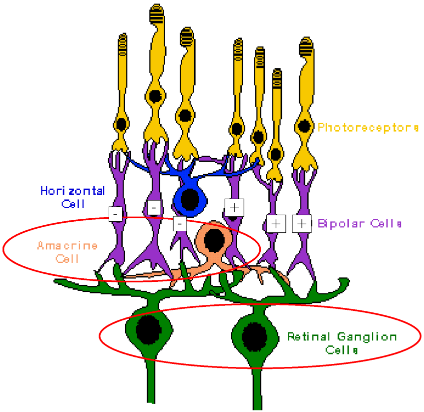
- The center-surround operation can be approximated by a Difference of Gaussian (DoG) operator

Organization of the bipolar cells

On-center and off-center configurations

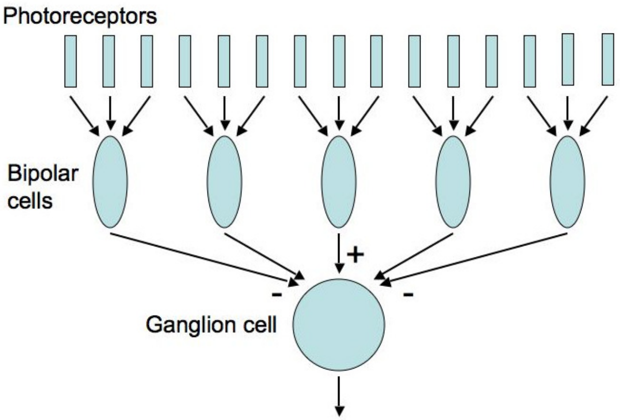


The amarcine and the ganglion cells



- Approx 126 million photosensors converge to approx 1 million optic nerves
  - Data reduction
- Amacrine cells contributes to motion detection

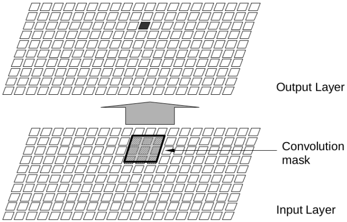
Model of receptive field



Generic model of signal processing in early vision

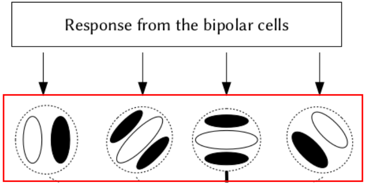
Digital Convolution (2D)

- Inputs:
  - Image $I = \{I_{xy}\}\ x = 1 : W, y = 1 : H$
  - Filter $F = \{F_{xy}\}\ x, y = -m : +m$
    - $[m \ll W, H]$
- Output: $I' = F * I = \{I'_{xy}\}$
  - $x = 1 : W, y = 1 : H$
  - where $I'_{xy} = \sum_{i=-m}^{m} \sum_{j=-m}^{m} F_{x+i,y+i}.I_{x-i,y-i}$



- Convolution is followed by a pooling layer for data reduction in CNN

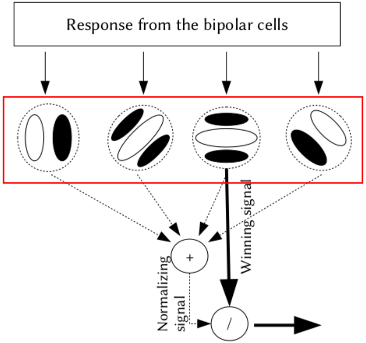
Oriented Filter Banks



- Center-surround organization of bipolar cells act like high-pass filters
- Configurations of the bipolar cells act as directional filters
  - Filters are differently oriented
  - Filters can be symmetric or asymmetric
- **Enables edge-detection in various directions**

Winner Take All (WTA) and Automatic Gain Control (AGC)

Applicable to all sensory signals



- The output of the filter with strongest output is transmitted
  - The strongest oriented edge is detected
- Output is normalized by the average response
  - Results in sublinear (logarithmic) perceptual response to a signal
  - Stronger signals are attenuated

Weber-Fechner Law

Holds good for all types of sensory signals

- Fechner's Law: Subjective sensation is proportional to the logarithm of the stimulus intensity
  - $P = K.\ln S + C$, **or**
  - $P = K.\ln(\frac{S}{S_0})$
- Weber's Law: The smallest change in stimuli that can be perceived is proportional to current signal strength
  - $\Delta S \propto S$

Weber-Fechner Law

Weber's Law can be derived from Fechner's law

- Fechner's Law: $P = K.\ln S + C$
- Differentiating:
  - $\frac{\Delta P}{\Delta S} = K.\frac{1}{S}$, **or**
  - $\Delta S = \frac{\Delta P}{K}.S$
- If $\tau = \Delta P$: the minimum change in perception that can be perceived,
  - The minimum perceivable change in stimulus $\Delta S = \frac{\tau}{K}.S$

Transformation in the eye

Retinal image

Neural image

- Human vision is sensitive to contrast, and not brightness
- There is huge data reduction in the early vision stage

Quiz

Quiz 02-03

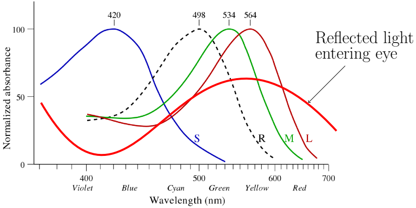
End of Module 02-03

IIT Jodhpur

Biological Vision and Applications
Module 02-04: Color Perception

Hiranmay Ghosh

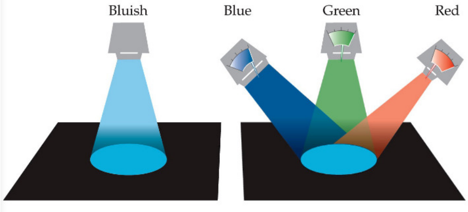
Cones and Color perception



- Incident light is characterized by $I(\lambda)$
- Let the response curve for the cones be $S(\lambda), M(\lambda)$ and $L(\lambda)$
- Exitation level of the S-cones is given by $E_S = \int_\lambda S(\lambda).I(\lambda).d\lambda$
  - ... and similarly for M- and L- cones
- Perceived color $C = f(E_S, E_M, E_L)$
  - Incident light of different spectra may result in the same color perception

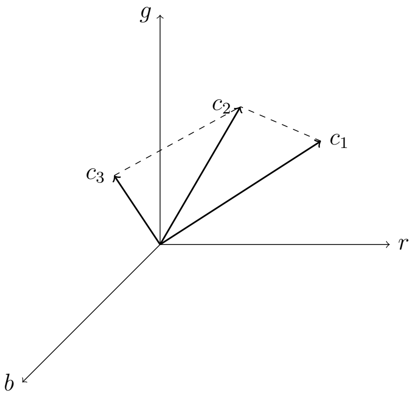
Trichromatic Color Theory

- Perceived color is a linear function of three independent variables
  - Response levels of the cones
- A perceived color can be matched by a linear combination of three primary colors
  - Proved by psychological experiments



Device dependent color models

RGB Model

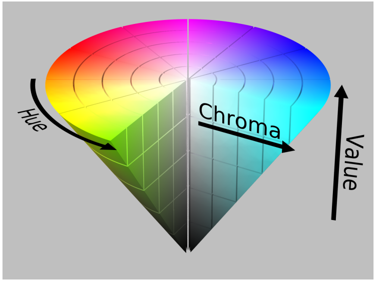


- Electronic devices typically use combination of red, green and blue to produce color
  - Combinations to be used depend on hardware characteristics
  - Device-dependent color model
- Each color is represented by a point in 3D space
  - Let $\vec{c_1}$, $\vec{c_2}$ and $\vec{c_3}$ represent three colors in *rgb* space
  - $\mid \vec{c_1} - \vec{c_2} \mid < \mid \vec{c_2} - \vec{c_3} \mid$ does not necessarily mean that
    - $\vec{c_2}$ is perceptually closer to $\vec{c_1}$ than $\vec{c_3}$

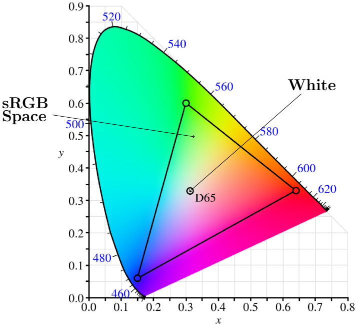
Device independent color models

HSV Model, CIE Model

- Munsell described color in terms of its three perceptual properties, namely
  - Hue (shade), Value (lightness), and Chroma (color purity)
- This is referred to as device-independent color model
- It has been later refined to many other models
  - HSV (Hue-Saturation-Value), CIE-XYZ and CIE-LAB
- In these models too, a color is represented by a point in a 3D space
  - The color distances in these spaces closely conform to perceptual distances



sRGB Color space



- Perceived color can be matched by a "linear combination" of three primary colors
  - The combination can involve addition and subtraction
- Unfortunately, we can only add (not subtract) color in electronic devices
  - We can produce only a subset of perceivable colors with the devices
- The color space that can be produced by a device is called sRGB space
  - Depends of the device characteristics

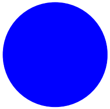
Are 24-bits sufficient to represent all perceivable colors?

- Human eye can distinguish between
  - Approximately 128 different hues
  - Around 20 to 30 different saturation values (for each hue)
  - Between 60 and 100 different brightness levels
- Combinatorially, human eye can distinguish between roughly 300,000 – 350,000 different colour shades
- 24 bits has a provision to represent 16 million colour shades!
  - The issue is how we intelligently utilize the 24 bits

Opponent process theory

Experiment

- Concentrate on the blue cicle below for about 10 seconds and then shift your gaze to the white area of the screen



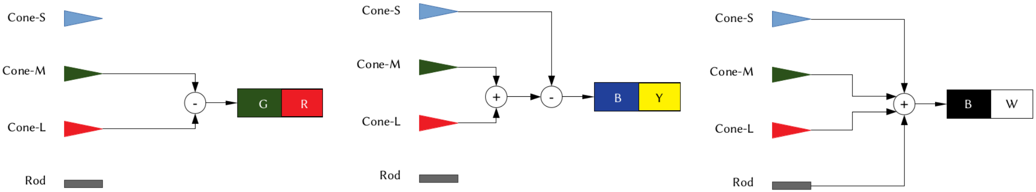
Opponent process theory

(Continued)

- You must have seen an yellow "after image" in the last slide
- The neural network connects the photo-receptors in a certain way to distinguish between three opponent color pairs
  - red *vs.* green
  - blue *vs.* yellow
  - dark (black) *vs.* bright (white)

Opponent process theory

(Continued)



Further reading

- An excellent blog on color science
  - https://medium.com/hipster-color-science/a-beginners-guide-to-colorimetry-401f1830b65a

Quiz

Quiz 02-04

End of Module 02-04

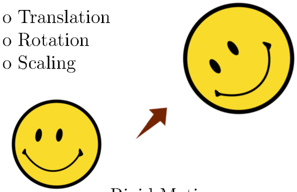
Biological Vision and Applications
Module 02-05: Motion Perception

Hiranmay Ghosh

Rigid, Elastic and Fluid Motion

- **Rigid motion** is where the moving object does not change shape
- **Elastic motion** is where the moving object changes shape with some continuity
- **Fluid motion** is where the continuity is not there

o Translation
o Rotation
o Scaling



Rigid Motion

Elastic Motion

What is the problem?

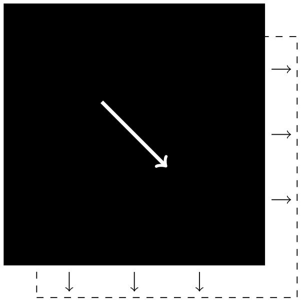
- An image is motion: $I(x, y, t)$
- The motion: $\vec{V}(x, y, t)$
- How to estimate $\vec{V}(x, y, t)$ from values of $I(x, y, t)$ observed over time
- Sometimes it is sufficient to detect motion
    - Measurement not necessary

Continuous and Discrete Motion

- Human observers can distinguish two types of motion
  - Continuous
  - Discrete
- To recognize continuous motion, an object need not move continuously over retinal field
  - Examples: Alternately blinking on of festive decoration lights, movie / TV
- There are two stages of motion recognition
  - Short range (60 - 100ms, 10 - 15' of visual arc): Based on local intensity changes
    - Local contrasts: early vision
  - Long range (400ms): Based on token matching
    - Object recognition: late vision

Intensity based scheme

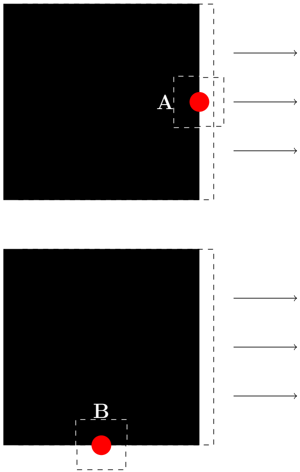
Basic Principle



- See the contour changing and infer motion
- Works well when there is significant intensity variation
  - Can be applied to object boundaries if there is a significant contrast between FG and BG

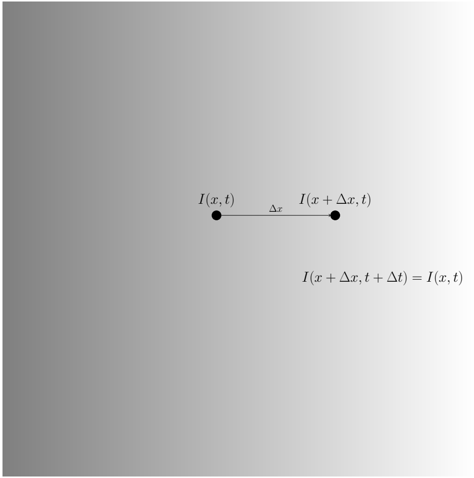
Estimating local motion

Aperture problem



- The motion can be perceived near point A
  - Intensity changes with time
- The motion cannot be perceived near point B
  - No intensity changes with time
  - Aperture problem

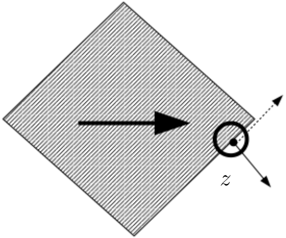
Gradient model of motion estimation



- $\Delta I = I(x+\Delta x, t) - I(x,t)$
- $\Delta x = \Delta I / \frac{\partial I}{\partial x}$
- $\Delta t = -\Delta I / \frac{\partial I}{\partial t}$
- $v = \frac{\Delta x}{\Delta t} = -\frac{\partial I/\partial t}{\partial I/\partial x}$

Gradient model of motion estimation

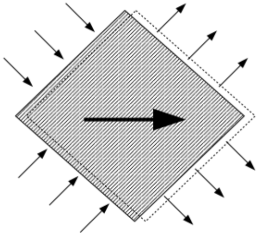
contd.



- Motion is estimated from the local gradients of the image intensity.
- The local velocity at z, in the direction of the spatial intensity gradient
- $v(z,t)_{\nabla I} = -\frac{I_t(z,t)}{|\nabla I(z,t)|}$
  - where
  - $I_t(z,t)$ represents the temporal gradient for local illumination change
  - $\mid \nabla I(z,t) \mid$ represents the magnitude of spatial gradient for local illumination change

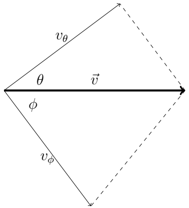
Rigid motion in image plane

Constant velocity assumption (translation only)



- The overall 2D motion of a rigid object can be estimated from the perceived motion at various points on the contour.

- Need minimum two points ($\theta \neq \phi$)
  - Let $\vec{V} = (v_x, v_y)$
  - $v_\theta = v_x.cos\theta + v_y.sin\theta$
  - $v_\phi = v_x.cos\phi + v_y.sin\phi$
  - Solve for v_x, v_y



Error resilience

Why we should observe at many points

- Minimize RMS error in

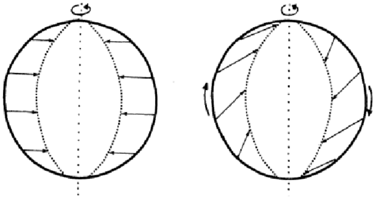
$$\begin{bmatrix} \alpha_1 & \beta_1 & -v_1 \\ \alpha_2 & \beta_2 & -v_2 \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ \alpha_n & \beta_n & -v_n \end{bmatrix} \begin{bmatrix} v_x \\ v_y \\ 1 \end{bmatrix} = 0$$

- where $\alpha_1 = cos\theta$, $\beta_1 = sin\theta$, $v_1 = v_\theta$ etc.
- Use Singular Value Decomposition (SVD)

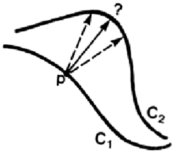
SVD through example

Ambiguity in motion estimation

More general cases



Motion in 3D

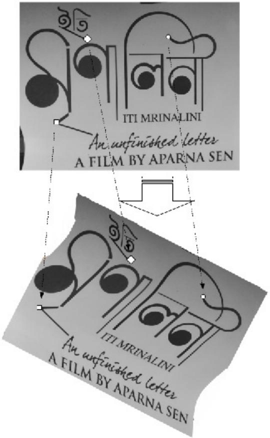


Elastic motion

- Sources of information loss
  - Projection of 3D object to 2D image
  - Projection of movement to intensity variation
- $\vec{V} = v_{\perp}.\vec{u}_{\perp} + v_{\top}.\vec{u}_{\top}$
  - $v_{\top}$ cannot be estimated
- Assumption on additional constraints are needed to estimate $v_{\top}$

Token based method

Motivated by higher level perception (token recognition)



- Tokens (distinctive points) are identified in the scene
  - Feature points (SIFT, SURF, etc.) can be used
- Tokens are tracked over time
  - Motion at tokens are estimated
  - Motion at other points interpolated

Token based method

(Continued)

- Depends of successful tracking of tokens
- Not an easy problem
  - Appearance of tokens may change
  - Two tokens are similar
- Tokens may be confused with each other during motion
- Additional domain-specific constraints need to be imposed
  - Relative geometry of tokens are maintained
  - Tokens have moved minimum distance
- Sometimes leads to illusion
  - A fan or a bicycle wheel appears to rotate in the opposite direction

Quiz

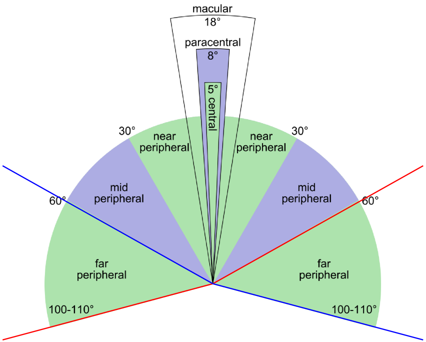
Quiz 02-05

End of Module 02-05

Biological Vision and Applications
Module 02-06: Peripheral vision

Hiranmay Ghosh

Foveal Vision and Peripheral Vision

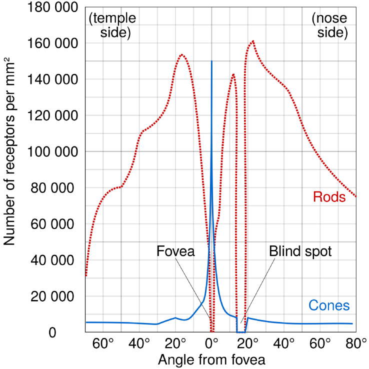


- Peripheral vision refers to vision beyond about 2 - 2.5^o from center of the eye
- Overlap area for both eyes is about 120^o
- Far peripheral region is seen with one eye
- Vision in mid/far peripheral region is predominantly black-and-white

Role of preipheral vision

- 99% of visual field is covered by peripheral vision
- Provides an approximate description of the visual field
- Useful for
  - Controls eye movement (saccade) in visual search
  - Shifts attention to desired place in image

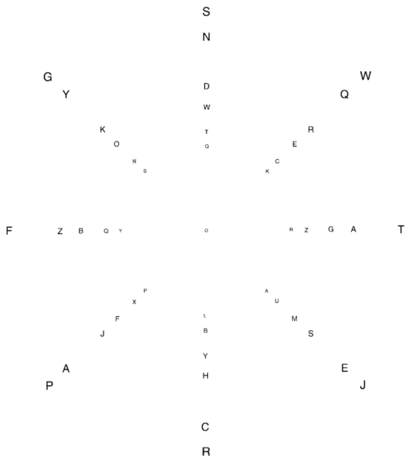

Cortical magnification



- As we move away from the foveal area of an eye
  - Linear decrease in rod density
  - The concentration of optic nerves also decreases.
  - rod:optic nerve ratio approx 600:1 at the far peripheral region
- Cortical magnification: equal volume of neurons cover more and more visual area
  - Less information is available

Effect of Cortical magnification

Minimum size of recognizable objects get bigger



Effect of Cortical magnification

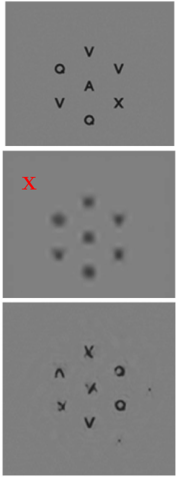
Crowding

- Focus on the cross-hair. Try to see the lettr ‘G’ in the left image, and in the right image

KGP + G

Model of Cortical magnification

Some distinctive textures are retained



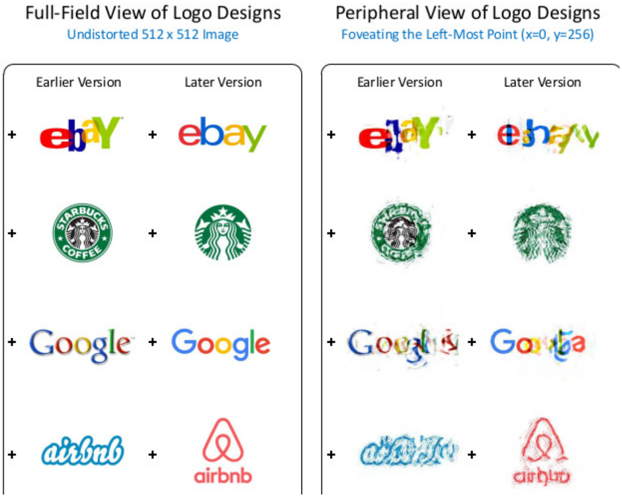
- Image compression (pooling / wavelet decomposition) results in blurred image
- Cortical magnification does not
  - Some distinctive textures are retained
  - There may be some disparity regarding locations
- The distinctive patterns help peripheral vision to guide the foveal vision in visual search
- Mathematical models for the peripheral texture representation
  - Summary statistics: autocorrelation and pooling
  - See Portilla 2000

Mongrels

- Mongrel: synthesized image to have the same summary statistics as a given original stimulus.
  - There can be many mongrels to an original stimulus

An interesting application

Logo design



Quiz

Quiz 02-06

End of Module 02-06

IIT Jodhpur

Biological Vision and Applications

Module 03-01: Reasoning

Hiranmay Ghosh

What is “reasoning”

- We “know” some facts
  - Supplied by others
  - Sensed by some sensors
- We infer unknown facts from the known facts

A simple example:

- Prior knowledge:
  - I need to go to the institute
  - Metro connects my home to the institute
- Inference:
  - Therefore, I take metro

Reasoning paradigms

- In human mind, reasoning is intuitive
- Formalizations
  - Knowledge driven (top-down)
    - Rule-based reasoning
    - Model-based reasoning
    - Case-based reasoning
  - Data driven (bottom-up)

Rule-based reasoning

- Apply some rules on the known facts to “deduce” unknown facts
- Formalized as logic
  - $\forall x : \mathit{bird}(x) \rightarrow \mathit{fly}(x), \forall x : \mathit{parrot}(x) \rightarrow \mathit{bird}(x)$
  - $\Longrightarrow \forall x : \mathit{parrot}(x) \rightarrow \mathit{fly}(x)$
- Also called “deductive” reasoning
- Many flavors
  - Propositional calculus, predicate calculus
  - First order logic, second order logic, ...
  - Descriptions logic
- See Norvig & Russell

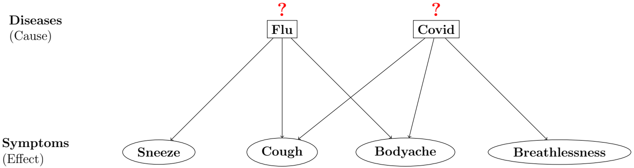
Rule-based reasoning

Strengths and weaknesses

- Major strength:
  - Reasoning is valid
  - If the premises are true, the consequence must be true.
    - Can be proved
- Major weaknesses:
  - Can discover fact implied by known facts
    - Cannot find "new" fact
  - Cannot handle uncertain sensory data
    - If premises are not known or incorrect, the reasoning breaks down

Model based reasoning

- Based on a mental model of the world (diseases and symptoms)
  - What symptoms are caused by a disease



- Doctor "observes" the symptoms (patient / lab reports)
- Doctor needs to infer the disease
  - Checks which model matches the observations

Why deductive reasoning cannot be used ?

- One could frame the rules
  - Sneeze $\wedge$ Cough $\wedge$ Body-ache $\rightarrow$ Flu
  - Cough $\wedge$ Body-ache $\wedge$ Breathlessness $\rightarrow$ Covid
- Why it does not work?
  - Uncertainty of effects: All symptoms for a disease may not appear in a patient
  - Noisy / incomplete data: The patient / lab reports may be incomplete or wrong
  - Degree of symptom: Symptoms may manifest in various degrees (strong/mild)

Matching a model

Exact match vs. approximate match

- Real-world is extremely complex – we lack knowledge
  - Creating an "exact" model of the real world is impossible
- A real-world scenario will seldom match a model
- Model that <u>best</u> explains the observations is accepted as the inference
  - There may be many different ways to define the "best" explanation
- Inferencing by best explanation is known as "abduction"

Abduction

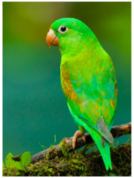
Strengths and weaknesses

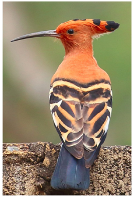
- Strengths
  - Robust against
    - Incomplete knowledge
    - Incomplete/erroneous data
  - Can generate “new” fact
    - Emergent knowledge
    - Observes symptoms – infers disease
- Weaknesses
  - The inference is not “valid”
    - Correctness of inference cannot be proven

Induction

Generalization from observations

- Example: Suppose you observe
  - Parrot is a bird; parrot can fly
  - Crow is a bird; crow can fly
  - Mynah is a bird; mynah can fly
  - ...
- Now we ask: Hoopoe is a bird; can it fly?
- From your earlier observations
  - You create a generalized model of a bird
  - You extrapolate the properties to a new species of bird
- You may get it wrong – penguins cannot fly




Properties of Abduction and Induction

- Abduction and induction result in new facts being generated
  - Not implied by existing knowledge
  - Inferred entities can be of different kind from the observed entities
- The inference of abduction and induction need not be valid
- Robust against uncertainties
  - In the model
  - In the observations
- Useful in processing sensory signals (essentially noisy)
- Induction is a special form of abduction

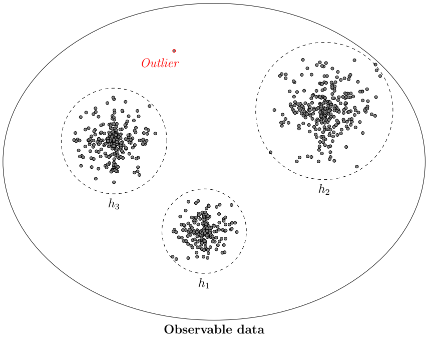
Case based reasoning

- Try to compare current scenario with scenario earlier observed
  - Infer based on earlier best-matched scenario
- Example
  - Earlier we have seen a winged thing – it could fly
  - We see a new winged thing – it should fly
- Apparently similar to induction
- Difference:
  - In induction, a generic model is formed (even without encountering a new scenario)
    - A new scenario is interpreted with the generic model
  - In CBR, no generic model is formed
    - A new scenario is compared with earlier cases
  - CBR can work with less experiential data

Data driven reasoning

Bottom-up approach

- Uses statistical similarity/associations to discover patterns
- We learn the models from data
- Flexible – no prior models
- Can't handle sparse and noisy data



Data driven reasoning

Example

	Sneeze	Cough	Body ache	Breathlessness
Patient 1	X	X	X	
Patient 2	X	X		
Patient 3		X	X	X
Patient 4		X	X	
Patient 5		X	X	X
Patient 6	X	X	X	
Patient 7		X		X
Patient 8	X		X	
Patient 9	X	X	X	
Patient 10			X	X

- No prior knowledge about diseases
- Patients 1, 2, 6, and 8 have similar symptoms $\rightarrow$ disease 1
- Patients 3,4,5,9 and 10 have similar symptoms $\rightarrow$ disease 2
- Patient 7 has Unique symptom
  - Observation error?
  - A new unknown disease?

- Pros: can discover new patterns (new models)
- Cons: inductive generalization not possible

Which one ?

- Which form of reasoning in used in the human mental processes ?
  - Probably all of them, depending on context
- Which form of reasoning in used in the human perception ?
  - Involves processing of sensory data (noisy)
  - Differences in visual appearance of object instances (uncertainties)
  - Incomplete model of the world (incomplete knowledge)
  - Model based abduction seems to be most appropriate
    - Inexact matching
    - Bayesian reasoning

Quiz

Quiz 03-01

End of Module 03-01

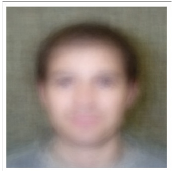
IIT Jodhpur

Biological Vision and Applications
Module 03-02: Model based reasoning for vision

Hiranmay Ghosh

Diversity in the natural world

- Each human face is different: makes modeling difficult



- But they exihibit some statistical similarity
- Super-imposition of several natural images
  - scale and pose normalized
  - Does not result in a blur background
  - Some statistical features stand out

Oliva and Torralba. The role of context in object recognition.

Natural scenes and human vision

- Natural scenes: images captured with devices operating in the range of visual spectrum.
  - Includes scenes of natural and man-made objects
  - Does not include text images, computer graphics, animations, paintings, cartoons, X-ray images, etc.

- Natural scenes are characterized by strong statistical regularities.
- Human eyes have adapted to the statistics of the natural scenes during the course of evolution
  - This has been the key to robust vision despite noisy image data

Vision as statistical interpretation

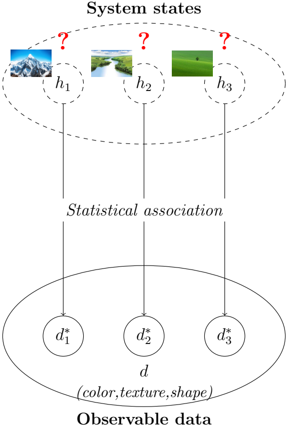
- We can "intuitively" reconstruct the occluded contour of the flower
- The statistical regularity is exploited to model vision as a process of statistical interpretation
  - Robust to natural variations / imperfections

Justification of feature based representation in Computer Vision

- A natural image can be represented as a point in $w \times h \times d$ dimensional space
  - w, h: width and height of the image
  - d: number of possible color values, e.g. $2^8 = 256$ for gray-scale images
- Combinatorially, it is possible to have $w \times h \times d$ distinct natural images of size $w \times h$
- Because of statistical regularity, most of the combinations never manifest
  - Natural images are clustered in very narrow regions in the image space
  - $\Rightarrow$ Lots of redundancies in the $w \times h \times d$ representation
  - $\Rightarrow$ Scope for compression
- A "feature" is an abstraction that characterizes the image contents
  - $\Rightarrow$ Lower dimensional representation of an image
  - Data compression

Model based reasoning for vision

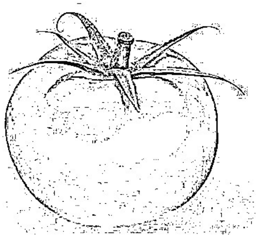
- The system is in one of the states (semantics)
  - $\mathcal{H} = \{h_1, h_2, \ldots h_n\}$
  - ... cannot be directly sensed
- A system state manifests itself in some observable (measurable) data d
  - Shape, color, texture
- Match data with expected manifestations
- Use statistical modeling and approximate match



$d \approx d_i^*$?, $d \approx d_2^*$?, $d \approx d_3^*$?

Early Vision and Scene Recognition

- An image is characterized by continuous homogeneous areas with interspersed discontinuities
  - Signify object contours in the scene
- Early vision detects the discontinuities (accentuates the contrasts)
  - Contour fragments are recognized
  - Noisy: Discontinuities / spurious edges
- Statistical properties of the contour fragments distribution leads to scene understanding



Quiz

No quiz for module 03-02

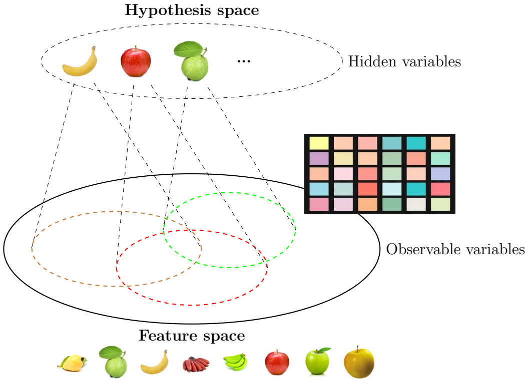
End of Module 03-02

IIT Jodhpur

Biological Vision and Applications
Module 03-03: Bayesian Reasoning for Vision

Hiranmay Ghosh

Vision is an inverted problem



- We do not see an object
  - We see the image features caused by the objects
  - We try to find best explanation for the observed image features
- This is an example of abductive reasoning
  - Bayesian reasoning is a probabilistic formulation

Baye's Theorem and Inferencing

Recap

Baye's Theorem:

$$P(A = a_i \mid B = b_j) = \frac{P(B=b_j|A=a_i).P(A=a_i)}{P(B=b_j)}$$

$$P(A \mid B) = \frac{1}{\kappa}.P(B \mid A).P(A)$$

where A and B are stochastic variables: $A = \{a_1, a_2, \ldots, a_m\}$, $B = \{b_1, b_2, \ldots, b_n\}$

- We try to infer the fruit from its color

Joint Probability Distribution

		Fruits (A)			
		Banana	**Apple**	**Guava**	**Total**
Color (B)	**Red**	0.07	0.1	0.01	*0.18*
	Green	0.21	0.04	0.07	*0.32*
	Yellow	0.42	0.06	0.02	*0.5*
	Total	*0.7*	*0.2*	*0.1*	*1*

$$P(banana \mid yellow) = \frac{0.42}{0.5} = 0.84$$

Using Baye's Theorem:

$$P(banana \mid yellow) = \frac{P(yellow|banana).P(banana)}{P(yellow)}$$
$$= \frac{\frac{0.42}{0.7} * 0.7}{0.5} = 0.84$$

Why Baye's Theorem ?

We do not have a complete knowledge about the world

		Fruits (A)				
		Banana	Apple	...	...	...
Color (B)	Red	0.1	0.5			
	Green	0.3	0.2			
	Yellow	0.6	0.3			
	Total	*1*	*1*			

$$P(Banana) = 0.7, P(Apple) = 0.2, P(Others) = 0.1$$

$$\underbrace{\boxed{P(banana \mid yellow)}}_{\color{red}{Posterior}} = \frac{1}{\kappa} * \underbrace{\boxed{P(yellow \mid banana).P(banana)}}_{\color{red}{Priors}}$$

$$= \frac{1}{\kappa} * 0.6 * 0.7 = \frac{1}{\kappa} * 0.42$$

How do we get to know the priors and κ ?

- Priors: $P(banana), P(yellow \mid banana)$
  - From external sources / context
  - From prior observations

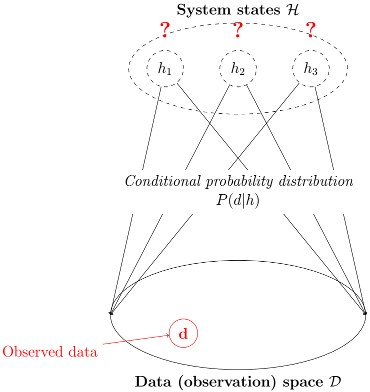
- Proportionality constant κ
  - We do not care
  - We need to find the best explanation
    - $P(banana \mid yellow) = \frac{1}{\kappa} * 0.42$
    - $P(apple \mid yellow) = \frac{1}{\kappa} * 0.06$
    - Banana is a better explanation than apple for observed yellow color

Bayesian Inference

Summary

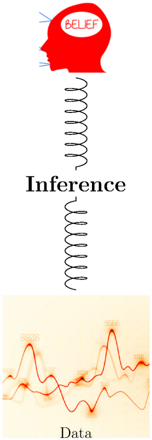
- Hypothesis space: $\mathcal{H} = \{h_1, h_2 \ldots h_m\}$
- Observable space: $\mathcal{D} = \{\mathbf{d}_1, \mathbf{d}_2 \ldots \mathbf{d}_n\}$
- Prior belief: $P(h_1), \ldots$
- Conditional probabilities: $P(\mathbf{d}_1 \mid h_1), \ldots$
- Observed data: $\mathbf{d} \in \mathcal{D}$

- Bayes formula: $P(h_i \mid \mathbf{d}) = \frac{P(\mathbf{d}|h_i).P(h_i)}{P(\mathbf{d})} = \frac{1}{\kappa}.P(\mathbf{d} \mid h_i).P(h_i)$
- Inference by best explanation (abduction):
  - $h^* = \text{argmax}_{h_i \in \mathcal{H}} P(h_i \mid \mathbf{d})$



Belief Revision

Prior belief and evidence



$$\textbf{Baye's Theorem: } P(h_i \mid \mathbf{d}) = \frac{1}{\kappa}.\underbrace{\boxed{P(h_i)}}_{\textit{Prior belief}}.\underbrace{\boxed{P(\mathbf{d} \mid h_i)}}_{\textit{Evidential support}}$$

- Bayesian inference is a synthesis of prior belief and evidence from observation
  - Key advantage over pure data-driven (machine learning) approach
- Challenge:
  - Strong prior belief: Takes lots of evidence to offset it
  - Weak prior belief: Susceptible to noisy data

Odds and log-Odds

$$\textbf{odds}(banana, apple \mid yellow) = \frac{P(banana|yellow)}{P(apple|yellow)}$$
$$= \frac{\frac{1}{\kappa} * 0.42}{\frac{1}{\kappa} * 0.06} = 7$$

$$\textbf{logodds}(banana, apple \mid yellow) = log \frac{P(banana|yellow)}{P(apple|yellow))}$$
$$= log(0.42) - log(0.06)$$
$$\approx (-0.38) - (-1.22) = 0.84$$

* log base assumed to be 10

- Useful for comparing the plausibility of pairs of concepts

Composite data

Data item $\mathbf{d}$ may be composite: $\mathbf{d} = (d_1, d_2, \ldots, d_n)$

		Fruits (A)				
		Banana	Apple	...	...	...
Color (B)	Red	0.1	0.5			
	Green	0.3	0.2			
	Yellow	0.6	0.3			
	Total	*1*	*1*			

		Fruits (A)				
		Banana	Apple	...	...	...
Shape	Long	0.8	0.3			
	Round	0.2	0.7			
	Total	*1*	*1*			

$$P(Banana) = 0.7, P(Apple) = 0.2, P(Others) = 0.1$$

- $\mathcal{D} = \{(Red, Long), (Red, Round), \ldots\}$
  - Conditionals: $P(Red, Long \mid Banana), \ldots$
  - Combinatorial explosion of data space makes modeling difficult
  - Data becomes sparse: there may be little data available for some rare combinations
- Assuming conditional independence of features

$$P(\mathbf{d} \mid h_i) = P(d_1 \mid h_i).P(d_2 \mid h_i).\ldots.P(d_n \mid h_i)$$

$$P(h_i \mid \mathbf{d}) = \frac{1}{\kappa}.P(h_i).\prod_{k=1}^{n} P(d_k \mid h_i)$$

$$\mathbf{logodds}(h_i, h_j \mid \mathbf{d}) = P(h_i) - P(h_j) + \sum_{k=1}^{n}(P(d_k \mid h_i) - P(d_k \mid h_j))$$

Advantages of modeling with Elementary data items

- Easier to model the statistical dependency of a hypothesis h_i with an elementary data item d_k than the composite d
  - Model size is additive, rather than combinatorial
  - Statistically more dependable
- Robust inference can be made with a subset of observations
  - Robust against missing / erroneous observations
  - Generally, it is possible to use a few discriminatory data elements

Example: Robust inference

- To recognize the object as a car, you need not consider all visual features of a car
  - Robust against occlusions, etc.

- You can reconstruct the contour of the occluded part of the image

- Can it be done with deductive reasoning?

Incremental belief update

- $P(h_i \mid \mathbf{d}) = \frac{1}{\kappa}.P(h_i).P(\mathbf{d} \mid h_i)$
- Assume that $\mathbf{d} = d_1, d_2, d_3, \ldots$ represents a data stream (possibly infinite)
- After d_1 arrives
  - Posterior: $P(h_i \mid d_1) = \frac{1}{\kappa_1}.P(h_i).P(d_1 \mid h_i)$
  - This posterior becomes the prior for the second observation
- After d_2 arrives
  - Posterior: $P(h_i \mid d_1, d_2) = \frac{1}{\kappa_2}.P(h_i \mid d_1).P(d_2 \mid h_i) = \frac{1}{\kappa_{12}}.P(h_i).P(d_1 \mid h_i).P(d_2 \mid h_i)$
  - This posterior becomes the prior for the third observation
- ... and so on
- System updates it's belief incrementally
  - Does sequence matter ?
- In practice, it may be possible to infer even before all data arrives

Example

		Fruits (A)				
		Banana	Apple	...	...	...
Color (B)	Red	0.1	0.5			
	Green	0.3	0.2			
	Yellow	0.6	0.3			
	Total	*1*	*1*			

		Fruits (A)				
		Banana	Apple	...	...	...
Shape	Long	0.8	0.3			
	Round	0.2	0.7			
	Total	*1*	*1*			

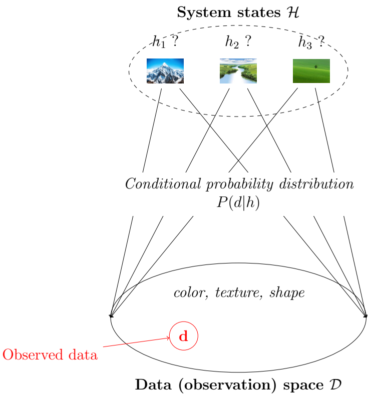
$$P(Banana) = 0.7, P(Apple) = 0.2, P(Others) = 0.1$$

- We see a green and long fruit.
  - Is it a banana or an apple ?
- Solution sketch:
  - Start with prior beliefs $P(banana)$ and $P(apple)$
  - "Observe" features in any order
    - Revise posterior beliefs for the fruits progressively
  - Check whichever is higher

> - Alternatively, use **odds**() / **logodds**()

Emergent knowledge

- We observe d
  - Visual patterns: color, texture, shape
- We infer h
  - Semantic concepts: mountain, river, greenery
- The inferred entities are of <u>different kind</u> than the observed entities
- New knowledge is created
- Paradigm applicable to higher layers of cognition also



Limitation of Bayesian reasoning

- We cannot infer an entity unless we have a model for it
  - The fruit was green and round. Was it really a guava ?
- A way to cope up for new concepts
  - Assume uniform probability distribution to begin with
  - Learn with experience
- Results are good only if
  - Prior belief is good
  - Model (conditionals) is good
  - Data (observation) is good
- Robust against imperfect priors / models / noisy data
  - We need best explanation, not accurate probability values

Does human mind follow Bayesian reasoning ?

- Both Yes & No

EdPuzzle: Cognitive bias

Quiz

Quiz 03-03

End of Module 03-03

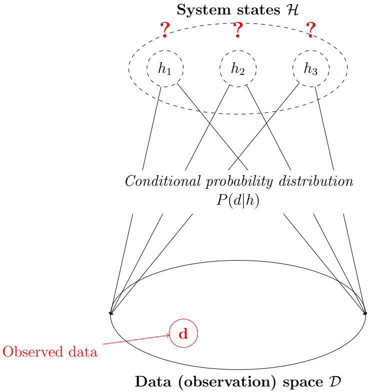
Biological Vision and Applications
Module 03-04: More on Bayesian reasoning

Hiranmay Ghosh

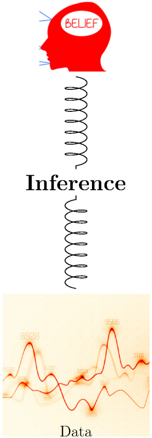
Bayesian Inference

Summary

- Hypothesis space: $\mathcal{H} = \{h_1, h_2 \dots h_m\}$
- Observable space: $\mathcal{D} = \{d_1, d_2 \dots d_n\}$
- Prior knowledge:
  - Prior probabilities: $P(h_1), P(h_2), \dots P(h_m)$
  - Conditional probabilities: $P(d_1 \mid h_1), P(d_2 \mid h_1) \dots P(d_n \mid h_m)$
- Observed data: $\mathrm{d} \in \mathcal{D}$
- Bayes formula: $P(h_i \mid \mathrm{d}) = \frac{P(\mathrm{d}|h_i).P(h_i)}{P(\mathrm{d})} = \frac{1}{\kappa}.P(\mathrm{d} \mid h_i).P(h_i)$
- Inference by best explanation (abduction):
  - $h^* = \mathrm{argmax}_{h_i \in \mathcal{H}} P(h_i \mid \mathrm{d})$



Prior knowledge and evidence



- Bayesian formula: $P(h_i \mid \mathrm{d}) = \frac{1}{\kappa}.P(\mathrm{d} \mid h_i).P(h_i)$
  - $P(h_i)$ represents prior belief in h_i
  - $P(\mathrm{d} \mid h_i)$ represents the evidential strength in support of h_i
  - Posterior belief $P(h_i \mid \mathrm{d})$ is the product of the two terms
- Bayesian inference is a synthesis of prior knowledge and evidence from observation
  - Key advantage over pure data-driven (machine learning) approach
  - Strong prior belief: Takes lots of evidence to offset it
  - Weak prior belief: Susceptible to noisy data

Odds

- Compare two hypotheses to find which one is more likely than the other
  - $\text{Odds}(h_i, h_j \mid d) = \frac{P(h_i|d)}{P(h_j|d)} = \frac{P(h_i)}{P(h_j)} \cdot \frac{P(d|h_i)}{P(d|h_j)}$
- Operating in logarithmic space makes the model additive
  - $\text{logOdds}(h_i, h_j \mid d) = (\log P(h_i) - \log P(h_j)) + (\log P(d \mid h_i) - \log P(d \mid h_j))$
- h_i is more likely than h_j iff $\text{Odds}(h_i, h_j \mid d) > 1$ or $\text{logOdds}(h_i, h_j \mid d) > 0$

$P(Banana) = 0.8, P(Apple) = 0.2$

		Fruits (A)				
		Banana	Apple	...	...	..
Color (B)	Red	0.1	0.6			
	Green	0.4	0.2			
	Yellow	0.5	0.2			
	Total	*1*	*1*			

- $P(B|Y) = 0.4 \times k$
- $P(A|Y) = 0.04 \times k$
- $\text{Odds}(B, A \mid Y) = \frac{0.4 \times k}{0.04 \times k} = 10$
- $\text{logOdds}(B, A \mid Y) = \log 10 = 1$

Composite data

- Data item d may be composite: $d = (d_1, d_2, \ldots, d_n)$
  - e.g. color, texture, shape
- Combinatorial explosion of data space makes modeling difficult
- Assuming conditional independence
  - $P(d \mid h) = P(d_1 \mid h).P(d_2 \mid h). \ldots P(d_n \mid h)$
- $P(h_i \mid d) = k.P(h_i). \prod_{k=1}^{n} P(d_k \mid h)$

$$\boxed{\begin{array}{l} \mathrm{Odds}(h_i, h_j \mid d) = \frac{P(h_i)}{P(h_j)} \times \prod_{k=1}^{n} \frac{P(d_k|h_i)}{P(d_k|h_j)} \\[2em] \mathrm{logOdds}(h_i, h_j \mid d) = (P(h_i) - P(h_j)) + \sum_{k=1}^{n} (P(d_k \mid h_i) - P(d_k \mid h_j)) \end{array}}$$

Advantages of modeling with Elementary data items

- Easier to model the statistical dependency of a hypothesis h_i with an elementary data item d_k than the composite d
  - The data space combinatorially expands with number of elementary items
  - Data becomes sparse – there may not be any data available for some rare combinations
- Robust inference can be made with a subset of observations
  - Robust against missing observations
  - Generally, it is possible to use a few discriminatory data elements
  - Wrong observations have less impact
  - Incremental belief update

Example: Robust inference

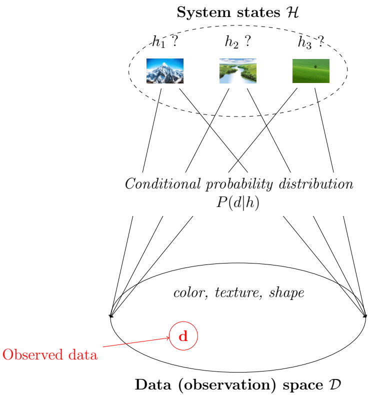
- To recognize the object as a car, you need not consider all visual features of a car
  - Robust against occlusions, etc.
- Can it be done with deductive reasoning?

Incremental belief update

- $P(h_i \mid d) = k.P(h_i).P(d \mid h_i)$
- Assume that $d = d_1, d_2 \ldots d_k$ represents a data stream
- After d_1 arrives
  - Posterior: $P(h_i \mid d_1) = k_1.P(h_i).P(d_1 \mid h_i)$
  - This posterior becomes the prior for the second observation
- After d_2 arrives
  - Posterior: $P(h_i \mid d_1, d_2) = k_2.P(h_i \mid d_1).P(d_2 \mid h_i) = k_{12}.P(h_i).P(d_1 \mid h_i).P(d_2 \mid h_i)$
  - This posterior becomes the prior for the third observation
- ...
- System updates it's belief incrementally
- In practice, it may be possible to infer even before all data arrives

Emergent knowledge

- We observe *d*
  - Visual patterns: color, texture, shape
- We infer *h*
  - Semantic concepts: mountain, river, greenery
- The inferred entities are of different kind than the observed entities
- New knowledge is created
- Paradigm applicable to higher layers of cognition also



Limitation of Bayesian reasoning

- We cannot infer an entity unless we have a model for it
  - Was that fruit really a kiwi ?
  - A way to cope up
    - Assume uniform probability distribution (0.33 for each color) to begin with
    - Learn (update probabilities) with experience
- Results are good only if
  - Model (priors, conditionals) is good
  - Data (observation) is good
  - Robust against imperfect models / noisy data

Quiz

Quiz 03-04

End of Module 03-04

Biological Vision and Applications
Module 03-05: Conditional independence

Hiranmay Ghosh

Joint probability

Recap

$$N = 80$$

	a_1	a_2
b_1	25	10
b_2	5	40

- Joint probability:
  $P(a_1 b_1) = \frac{|a_1 b_1|}{N} = \frac{25}{80} \approx 0.31$
- Marginal probability:
  $P(a_1) = \frac{|a_1|}{N} = \frac{30}{80} \approx 0.38$
  - $|\, a_1 \,| = |\, a_1 b_1 \,| + |\, a_1 b_2 \,| = 25 + 5 = 30$
- Conditional probability:
  $P(a_1 \mid b_1) = \frac{|a_1 b_1|}{|b_1|} = \frac{25}{35} \approx 0.71$
  - $|\, b_1 \,| = |\, a_1 b_1 \,| + |\, a_2 b_1 \,| = 25 + 10 = 35$

Conditional Independence

- The variable A is conditionally independent of B, iff
  - the states of A does not depend on the states of B
- Formally $\forall i = 1..m, j = 1..n : P(a_i \mid b_j) = P(a_i)$
- Now $P(a_i \mid b_j) = \frac{P(a_i b_j)}{P(a_j)}$
- Substituting, condition for "conditional independence" in symmetric form
  - $\forall i, j : P(a_i b_j) = P(a_i).P(b_j)$
- Conditional independence is symmetric
  - $CondInd(A, B) \Leftrightarrow CondInd(B, A)$

Exercise: Prove that if $\forall i, j : P(a_i \mid b_j) = P(a_i)$, then $\forall i, j : P(b_j \mid a_i) = P(b_j)$

Conditional Independence

Continued

$$N = 80$$

	a_1	a_2
b_1	25	10
b_2	5	40

- Are the variables A and B conditionally independent ?
  - $P(a_1) = \frac{30}{80} \approx 0.38$
  - $P(a_1 \mid b_1) = \frac{25}{35} \approx 0.71$
  - $P(a_1) \neq P(a_1 \mid b_1)$
    - A and B are not conditionally independent

Case of three variables

Place
o Tamil Nadu
o Kashmir

Fruit
o Banana
o Apple

Color
o Red
o Green
o Yellow

$P(B \mid T) = 0.8$
$P(A \mid T) = 0.2$

$P(B \mid K) = 0.1$
$P(A \mid K) = 0.9$

$P(R \mid B) = 0.1$
$P(G \mid B) = 0.4$
$P(Y \mid B) = 0.5$

$P(R \mid A) = 0.6$
$P(G \mid A) = 0.2$
$P(Y \mid A) = 0.2$

- Place not specified: $P(Y) = ?$
  - $P(B) = 0.7 \times 0.8 + 0.3 \times 0.1 = 0.59$
  - $P(A) = 0.7 \times 0.2 + 0.3 \times 0.9 = 0.41$
  - $P(Y) = 0.59 \times 0.5 + 0.41 \times 0.2 = 0.377$
- Place specified as Kashmir: $P(Y \mid K) = ?$
  - $P(B) = 0.1$
  - $P(A) = 0.9$
  - $P(Y) = 0.1 \times 0.5 + 0.9 \times 0.2 = 0.23$
- $P(Y) \neq P(Y \mid K)$
  - Place and Color are not conditionally independent

Case of three variables

contd.

Place
- o Tamil Nadu
- o Kashmir

$P(B \mid T) = 0.8$
$P(A \mid T) = 0.2$

$P(B \mid K) = 0.1$
$P(A \mid K) = 0.9$

Fruit
- o Banana
- o Apple

Color
- o Red
- o Green
- o Yellow

$P(R \mid B) = 0.1$
$P(G \mid B) = 0.4$
$P(Y \mid B) = 0.5$

$P(R \mid A) = 0.6$
$P(G \mid A) = 0.2$
$P(Y \mid A) = 0.2$

- Given that Fruit = Banana
- Place not specified
  - $P(Y) = P(Y \mid B) = 0.5$
- Place specified as Kashmir:
  - $P(Y \mid K, B) = P(Y \mid B) = 0.5$
  - ... similarly, for TN, for other colors, for apple
- Given F, P and C are conditionally independent

Quiz

Quiz 03-05

End of Module 03-05

Biological Vision and Applications
Module 03-06: Bayesian network

Hiranmay Ghosh

Bayesian reasoning revisited

- Bayesian framework of reasoning
  - Create a system model in terms of n stochastic (random) variables
    - $\mathcal{X} = \{X_1, X_2, X_3, \ldots, X_n\}$
    - A variable X_i can have k_i states. $X_i : \{x_i^1, x_i^2, \ldots x_i^{k_i}\}$
  - Some variables are observable, some are hidden (to be inferred)
  - Inference is a result of probability updates based on the observed data
- The joint probability distribution table will contain $\prod_i k_i - 1$ **independent** entries
- That is a big number !
  - A trivial system with 10 binary variables will have $2^{10} - 1 = 1023$ entries

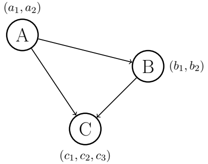
Joint probability and conditional probability

Joint probabilities

$P(a_1, b_1, c_1)$	$P(a_2, b_1, c_1)$
$P(a_1, b_1, c_2)$	$P(a_2, b_1, c_2)$
$P(a_1, b_1, c_3)$	$P(a_2, b_1, c_3)$
$P(a_1, b_2, c_1)$	$P(a_2, b_2, c_1)$
$P(a_1, b_2, c_2)$	$P(a_2, b_2, c_2)$
$P(a_1, b_2, c_3)$	

Conditional probabilities

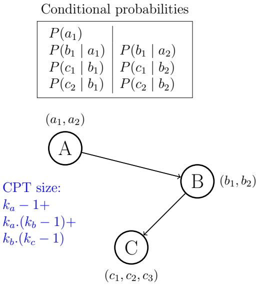
$P(a_1)$	
$P(b_1 \mid a_1)$	$P(b_1 \mid a_2)$
$P(c_1 \mid a_1, b_1)$	$P(c_1 \mid a_2, b_1)$
$P(c_1 \mid a_1, b_2)$	$P(c_1 \mid a_2, b_2)$
$P(c_2 \mid a_1, b_1)$	$P(c_2 \mid a_2, b_1)$
$P(c_2 \mid a_1, b_2)$	$P(c_2 \mid a_2, b_2)$



Non-circular depencency between variables assumed

- Consider three variables
  - A: $\{a_1, a_2\}$, B: $\{b_1, b_2\}$, C: $\{c_1, c_2, c_3\}$
- Joint probability table will have 11 independent entries
- Equivalently, they can be expressed with 11 conditional probabilities
- The joint probabilities can be computed from the conditional probabilities, e.g.
  - $P(a_1, b_1) = P(a_1).P(b_1 \mid a_1)$
  - $P(a_1, b_1, c_1) = P(a_1, b_1).P(c_1 \mid a_1, b_1)$
    - $= P(a_1).P(b_1 \mid a_1).P(c_1 \mid a_1, b_1)$

Conditional Independence



Conditional independence between variables A and C assumed

- Many of the variables in real world are conditionally independent of each other
  - Color of a fruit and place
  - Outcome of tossing of two (biased/unbiased) coins
  - Symptoms of a disease (headache, fever, ...)
  - Visual features of an object (color, shape ...)
  - ...
- Conditional independence simplifies probability computations
  - Another reason to work with conditional probabilities

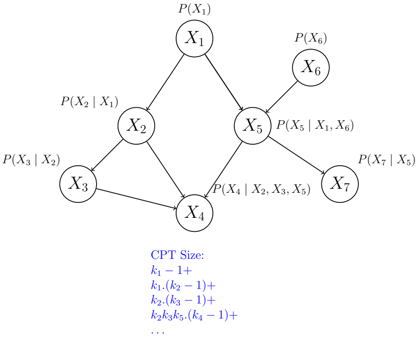
Probabilistic Graphical Models

- Graphical models exploit conditional independence
- The variables are depicted as nodes in the graph
- Only the variables that are not conditionally independent are connected with edges
- Generally a graph is sparse
  - Size of the CPT is much smaller than exhaustive joint distribution table
- There are many probabilistic graphical models
  - Markov Field, Hidden Markov Model, Bayesian Network, ...

See Koller. Probabilistic Graphical Models (book) / Cousera course

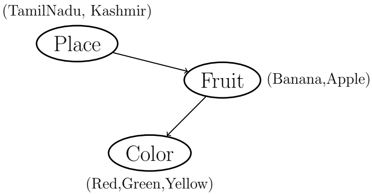
Bayesian Networks (BN)

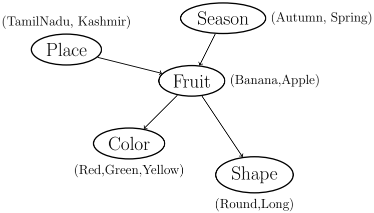
Models a probabilistic reasoning problem with cause-effect relationship



- A Directed Acyclic Graph (DAG)
- Nodes represent events in a system
  - $X_i = (x_i^1, x_i^2, \ldots, x_i^{k_i})$
  - Some nodes are observable; others need to be inferred
- Edges represent **causal** relations between the events
- Conditional probabilities $P(X_i \mid \mathrm{Pa}(X_i))$are associated with every node
  - where $\mathrm{Pa}(X_i)$ represents the parent set of node X_i

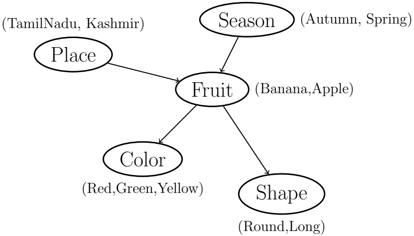
Examples of BN





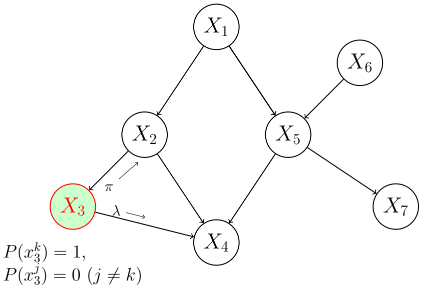
- You have already seen this example

Causal inference and Evidential inference



- Fruit is inferred from
  - Where you are, what is the season (Causal reasoning)
  - It's color and shape (Evidential reasoning)
- Bayesian network supports both types of reasoning

Inferencing with Bayesian Networks

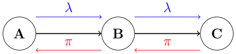


Pearl's algorithm

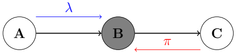
- Hand compute probabilities
  - There can be multiple (undirected) paths between a pair on nodes
  - Extremely complex
- Pearl's belief propagation algorithm
  - π and λ messages
    - Probabilities of neighboring nodes updated
  - Traverses recursively in the network
    - Till no more nodes left / blocked

Network topology and Belief propagation

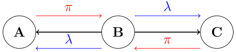
D-Separation



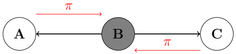
A causes **B**, **B** causes **C**, State of **B** is unknown
Belief flows between **A** and **C**



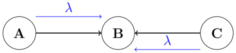
A causes **B**, **B** causes by **C**, State of **B** is known
The path between **A** and **C** is blocked by **B**



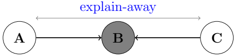
B is the cause of **A** and **C**, state of **B** is unknown
Belief flows between **A** and **C**



A and **C** are causes of **B**, state of **B** is known
The path between **A** and **C** is blocked by **B**



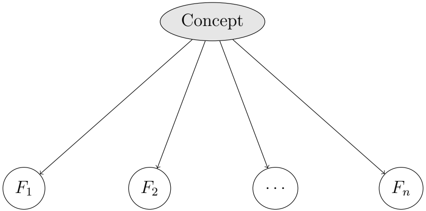
A and **C** are causes of **B**, state of **B** is unknown
A and **C** are conditionally independent



A and **C** are causes of **B**, state of **B** is known
A explains away **C**, and vice-versa

- Belief flows between two nodes in a network if there is an unblocked path between them
- If there are no unblocked path between two nodes, they are said to be <u>d-separated</u>

Naïve Bayesian Network

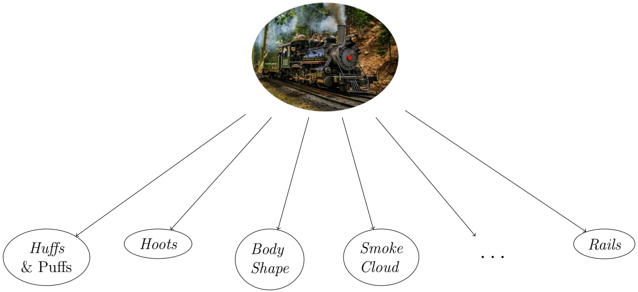


Features $F = (F_1, F_2, \ldots F_n)$

- Tree structure
- Two levels: concept and features
- Concept needs to be inferred from observed features
- Some times all features may not be observed

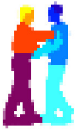
Naïve Bayesian Network

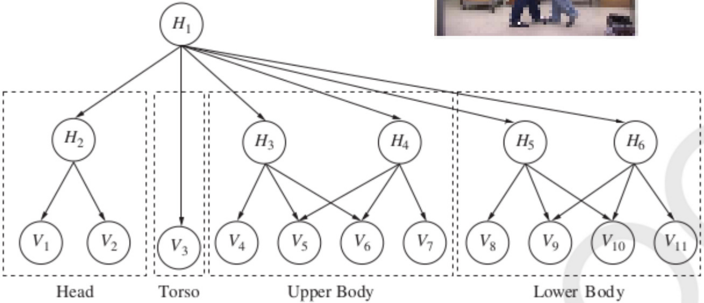
Example



Hierarchical organization in Bayesian Network

Example

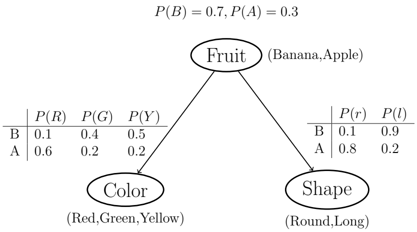




Park & Aggarwal. A hierarchical Bayesian network for event recognition of human actions and interactions

Simple Bayesian Network Example

Priors



- We need to evaluate the two hypotheses
  - Fruit is either Banana or Apple
- From the given data, we can find the prior marginal probabilities

Colors:

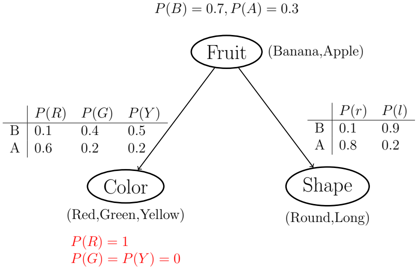
$$\begin{aligned}
P(Red) &= P(R \mid B) \times P(B) + P(R \mid A) \times P(A) \\
&= 0.1 \times 0.7 + 0.3 \times 0.6 = 0.25 \\
P(Green) &= 0.34 \\
P(Yellow) &= 0.41
\end{aligned}$$

Shapes:

$$\begin{aligned}
P(Round) &= 0.31 \\
P(Long) &= 0.69
\end{aligned}$$

Simple Bayesian Network Example

Posteriors



- We see a fruit to be red

Fruits: (un-normalized)

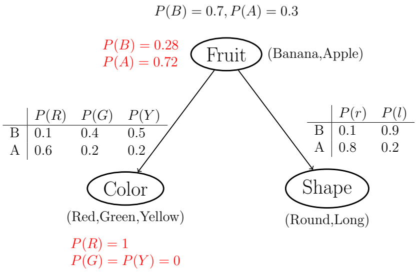
$$P(Banana \mid Red) = P(R \mid B) \times P(B)$$
$$= 0.1 \times 0.7 = 0.07$$
$$P(Apple \mid Red) = 0.18$$

Fruits: (normalized)

$$P(Banana \mid Red) = 0.28$$
$$P(Apple \mid Red) = 0.72$$

Simple Bayesian Network Example

Posteriors (contd.)



- Change in probability of fruits changes posterior probability of shapes

Shapes:

$$P(Round) = 0.60$$
$$P(Long) = 0.40$$

- Now we see the fruit to be **long**
  - Which fruit it is likely to be?

Quiz

Quiz 03-06

End of Module 03-06

Biological Vision and Applications
Module 03-07: Parameter Estimation

Hiranmay Ghosh

How to estimate a parameter ?

Maximum Likelihood estimation

- Bayesian framework of reasoning assumes some conditional probabilities (priors)
  - e.g., $P(\text{Red} \mid \text{Banana}) = 0.1$
- Where do you get the number from?
- Maximum likelihood estimation (purely data-driven):
  - You observe 20 bananas; 2 are red
  - $P(\text{red} \mid \text{banana}) = \frac{2}{20} = 0.1$
- Not reliable, if the sample size is small
  - Does not tell you how reliable the estimate is

Bayesian Theory provides a more reliable methods

Combines prior belief with observations

- Let the parameter $\theta = P(\text{red} \mid \text{banana})$
  - $\theta \in [0, 1]$
- **Prior hypotheses**: Without any further information, we may assume
  - All values of $\theta \in [0, 1]$ are equi-probable
  - i.e. the pdf $p(\theta)$ has a uniform distribution. $p(\theta) = 1$
- Now, we depend on data (observations) to update the belief
- By Baye's law

$$p(\theta \mid d) = \frac{P(d|\theta).p(\theta)}{P(d)}$$

$$\text{where } P(d) = \int_0^1 P(d \mid \theta).d\theta$$

$P(x)$ is probability (discrete), $p(x)$ is probability density function (continuous)

Bernoulli's Theorem

- Consider the problem
  - Suppose, you toss a coin t times
  - Probability of head is θ on every toss (known)
  - What is the probability of the outcome d: h heads and $t - h$ tails?
- Bernoulli's theorem:

$$P(d \mid \theta) = \theta^h.(1-\theta)^{t-h}$$

- Easy to derive. ... Try it out!

Back to our problem

Tossing a coin

- Using Bernoulli's theorem

$$\begin{aligned} P(d) &= \int_0^1 P(d \mid \theta).d(\theta) = \int_0^1 \theta^h.(1-\theta)^{t-h}.d\theta \\ &= \frac{h!.(t-h)!}{(t+1)!} \end{aligned} \tag{1}$$

- We have

$$p(\theta) = 1 \qquad \text{[Uniform probability assumption]} \tag{2}$$
$$P(d \mid \theta) = \theta^h.(1-\theta)^{t-h} \qquad \text{[Bernoulli's theorem]} \tag{3}$$
$$p(\theta \mid d) = \frac{P(d|\theta).p(\theta)}{P(d)} \qquad \text{[Bayes Theorem]} \tag{4}$$

- Substituting (1), (2), (3) in (4)

$$\boxed{p(\theta \mid d) = \frac{P(d|\theta).p(\theta)}{P(d)} = \frac{(t+1)!}{h!.(t-h)!}.\theta^h.(1-\theta)^{t-h}}$$

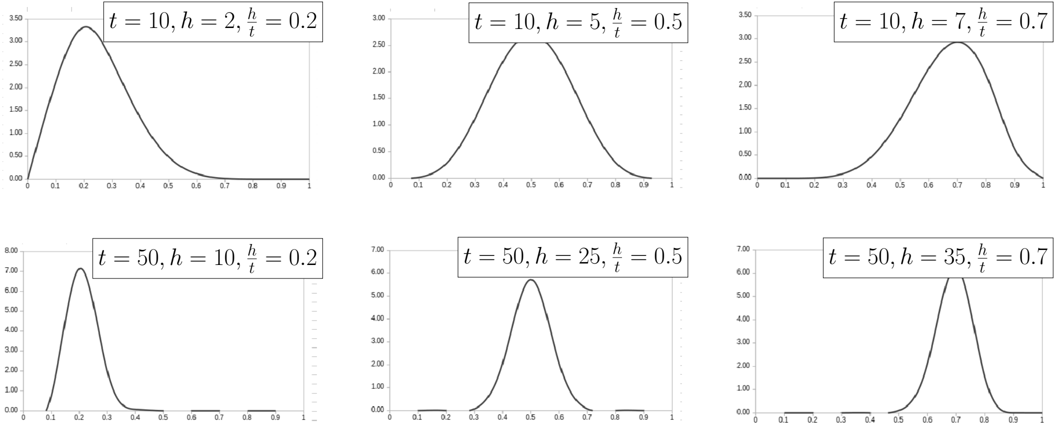
Discussions

- We get a pdf for θ, rather than a single value
  - More informative
- Expected value for θ is

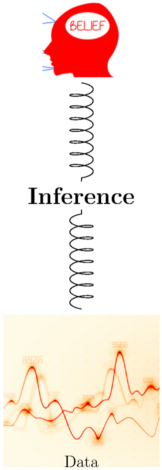
$$\boxed{\hat{\theta} = \int_0^1 \theta.p(\theta \mid d).d\theta = \frac{h+1}{t+2}}$$

- Bayesian method vs. max likelihood estimate
- Let's assume, we have observed 2 bananas, none is red
  - $h = 0, t = 2$
- By max. likelihood: $\theta = \frac{h}{t} = 0$
- By Bayesian method: $\hat{\theta} = \frac{h+1}{t+2} = \frac{1}{4}$
  - Prior belief in Bayesian method moderates the extreme estimates

Dependence of pdf on data



Priors vs. data (observations)



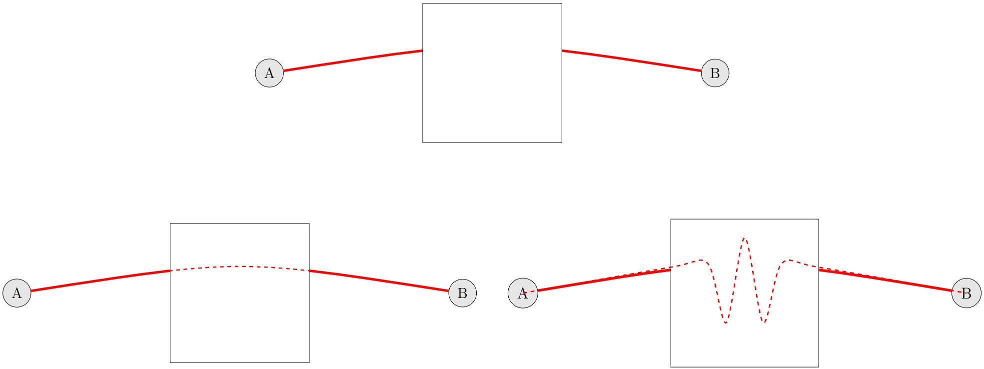
- We have assumed uniform pdf $p(\theta) = 1$ in this example.
  - It is possible of assume other priors
  - What determines the priors?

- Prior belief dominates so long there are less observations
- Data tends to dominate with increased number of observations

- Weak prior $\Rightarrow$ it takes less data to update the parameter
  - Susceptible to noisy data
- Strong prior $\Rightarrow$ it takes more data to update the parameter
  - Susceptible to erroneous prior

How do we get the priors ?

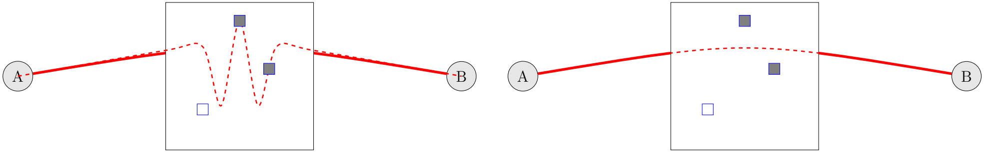
On complexity of models and priors



- Human mind tends to choose the simplest model

What do the data say?

Goodness of fit

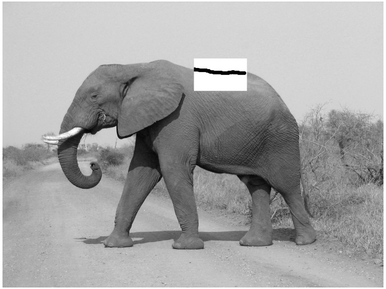


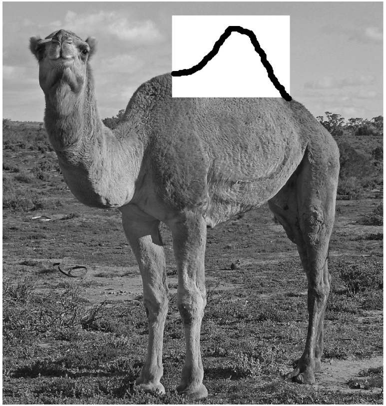
- How well does the fit a model?

Complexity and Prior

- Let $c(M)$ denote the complexity for a model M
  - Prior probability for M can be expressed as: $P(M) = 2^{-c(M)}$ (axiom)
  - Probabilies for the model hypotheses (priors and conditionals) and inference
    - $P(h_i) = 2^{-c(h_i)}$
    - $P(d \mid h_i) = 2^{-c(d|h_i)}$
    - $P(h_i \mid d) = 2^{-c(h_i|d)}$
- Baye's law $P(h_i \mid d) = \kappa.P(h_i).P(d \mid h_i)$
- Substituting, and taking logarithm
  - $c(h_i \mid d) = k + c(h_i) + c(d \mid h_i)$
- Human mind chooses the inference with least complexity
  - Complexity of inference is the sum of complexities of hypotheses
  - Belief maximization $\equiv$ complexity minimization

Complexity (prior probability) is guided by knowledge





Quiz

Quiz 03-07

End of Module 03-07

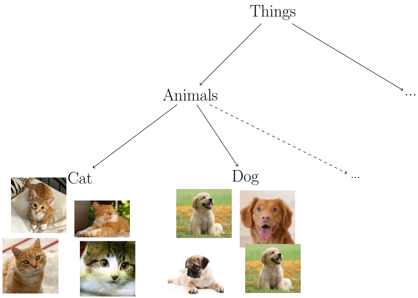
IIT Jodhpur

Biological Vision and Applications
Module 03-08: Taxonomy Learning

Hiranmay Ghosh

Taxonomy

Organizing concepts in a hierarchy

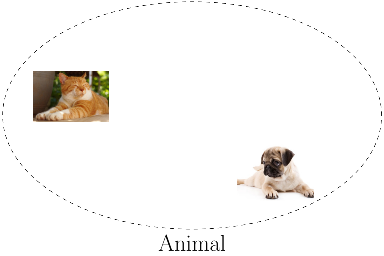


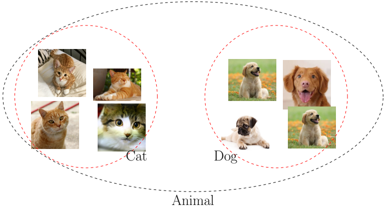
- Learned top-down, or bottom-up ?

Taxonomy Learning

The cognitive science viewpoint

- Psychologists suggest that it is learned top-down with experience





Taxonomy is a tradeoff

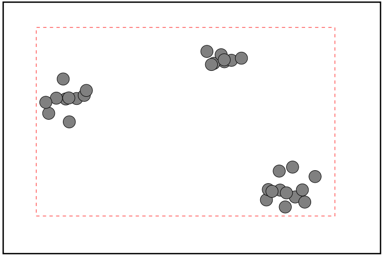
... between complexity of hypothesis and it's goodness of fit (with data)

- Complexity of hypothesis
  - Human mind accepts the simplest theory
  - A hypothesis with one class is simplest
    - ... more classes $\rightarrow$ more complexity
- "Goodness of fit"
  - Probability of data being explained by the hypothesis
  - Tighter the enclosed area, better is the goodness of fit

Example

Hypothesis 1

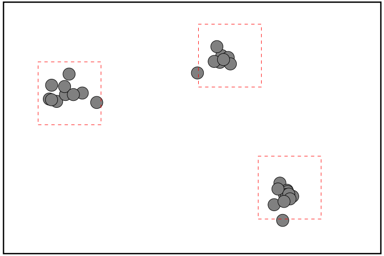
Simple (1 class)
Poor goodness of fit (large bounding box)



Feature space

Hypothesis 2

Complex (3 classes)
Better goodness of fit (small bounding boxes)



Feature space

- Which one is accepted ?

Bayesian approach for hypothesis selection

- Posterior probability of a hypothesis: $P(h_i \mid d) = k_1.P(h_i).P(d \mid h_i)$
  - Choose $h^* = \text{argmax}_i P(h_i \mid d)$
  - We can ignore k_1 (for comparisons)

- Prior for hypothesis: $P(h_i) = f(c_i)$
  - c_i: number of classes in hypothesis h_i (complexity)
  - $f(c_i)$: a exponentially decreasing function on c_i
    - $P(h_i) = k_2^{-c_i}$

Bayesian approach for hypothesis selection

... contd.

- "Goodness of fit": $P(d \mid h_i) = \left(\frac{k_3}{A}\right)^n$
  - Inverse of probability of the data to fit in the designated area (in feature space)
  - n: Total number of data points
  - A: Total area for all the category spaces
  - k_3: A constant (scaling factor for A)
- Posterior: $P(h \mid d) = k_1.k_2^{-c_i}.\left(\frac{k_3}{A}\right)^n$
- Generally ...
  - Simpler hypothesis dominates with sparse data
    - When we have seen less number of instances, we tend to believe that they are all in the same group
  - Goodness of fit dominates with dense data
    - When we have seen more instances, we tend to classify them in subgroups

Quiz

No quiz for module 03-08

End of Module 03-08

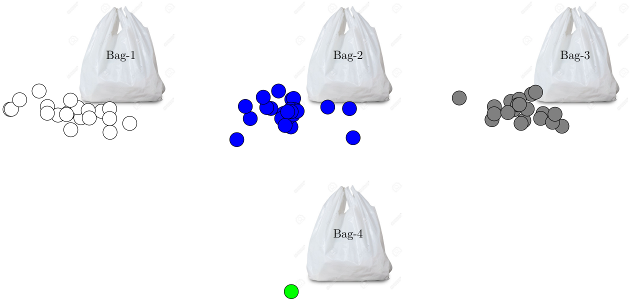
IIT Jodhpur

Biological Vision and Applications
Module 03-09: Hierarchical Bayesian Model

Hiranmay Ghosh

An example

The bags can have marbles



Specific knowledge and Generic knowledge

- Specific Knowledge
  - When we sample marbles from a particular bag, we gain knowledge about that bag
  - e.g. uniformity of colors and color of marbles in that bag
- Generic Knowledge
  - When we sample marbles from several bags, we gain knowledge about all bags
    - ... even those are not sampled
  - e.g. uniform color of marbles in each bag
- Specific knowledge about several bags lead to generic knowledge
  - This is an instance of inductive reasoning or inductive generalization
  - The process of gaining generic knowledge is also known as meta-learning

Modeling the problem

- Let $\vec{\theta}_i$ represent the model parameters for bag i
  - $\vec{\theta}_i = (\theta_{ij}, j = 1 \ldots n)$
  - j represents the different colors. $\sum_j \theta_{ij} = 1$
- In HBM
  - $\vec{\theta}_i$s are modeled as probabilistic functions of some hyper-parameters
  - The hyper-parameters represent a higher (more abstract) level of knowledge
- A common approach is to use Dirichlet distribution
  - In this example, parameters can be $\alpha, \vec{\beta}$
  - α represents the heterogeneity of colors of the marbles in the individual bags
  - $\vec{\beta}$ representing the average color distribution across all the bags

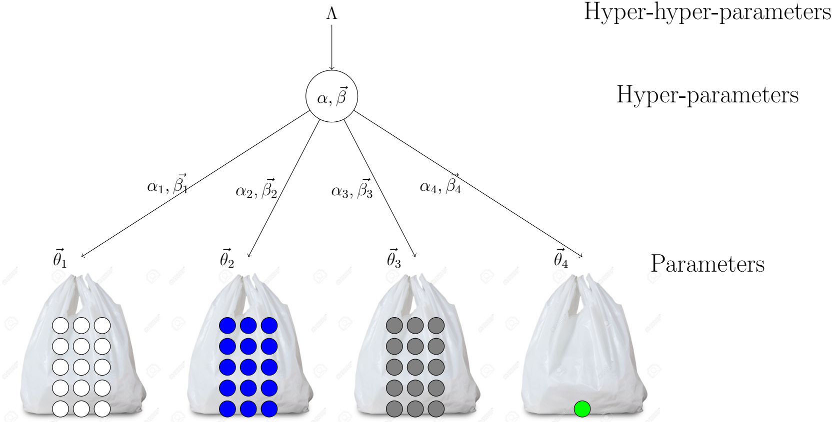
On Dirichlet distribution

- Beta Distribution
  - A probability distribution function with two parameters (α, β)
  - $p(\theta)_{\alpha,\beta} = \frac{1}{k}.\theta^{\alpha-1}.(1-\theta)^{\beta-1}$
  - where $k = \frac{\Gamma(\alpha).\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
  - where $\Gamma(x) = \int_{t=0}^{\infty} t^{x-1} e^{-t} dt$
- Dirichlet Distribution (a generalization of Beta distribution)
  - $\vec{\theta} = (\theta_1, \theta_2 \ldots \theta_n)$
  - $\vec{\alpha} = (\alpha_1, \alpha_2 \ldots \alpha_n)$
  - $p(\vec{\theta})_{\vec{\alpha}} = \frac{1}{k}.\prod_{i=1}^{n} \theta_i^{\alpha_i - 1}$
  - where $k = \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{n} \alpha_i)}$

https://leimao.github.io/blog/Introduction-to-Dirichlet-Distribution/

Hierarchical Bayesian Model

A graphical depiction



Discussions

- The models for the bags are linked with hyper-parameters $\alpha, \vec{\beta}$
  - Are learned together with θs
  - An observation for one bag serves as an observation for the other bags too
  - Hyper-parameters $\alpha, \vec{\beta}$ are learned together with the parameters θ_is

- $\vec{\theta}_i$ is a probabilistic function of $\alpha, \vec{\beta}$
  - α, $\vec{beta}$ impose constraints on values of $\vec{\theta}_i$s
  - Priors for $\vec{\theta}_i$s (no observations) are closer to actual values
  - $\vec{\theta}_i$ can be learned (reliably estimated) from less number of observations

- Hyper-parameters represent more abstract knowledge
- It is possible to model $\alpha, \vec{\beta}$ with even higher level of knowledge ...
  - Further inductive generalization is possible
  - Generalization from one problem to another will be efficient for similar problems

Quiz

Quiz 03-09

End of Module 03-09

Biological Vision and Applications
Module 03-10: Feature Learning

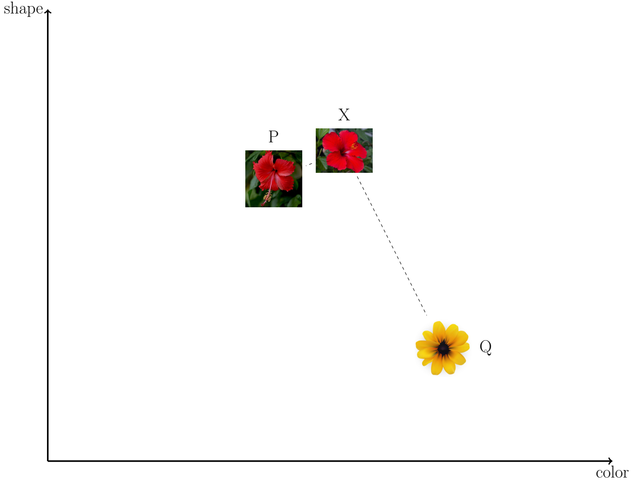
Hiranmay Ghosh

Which feature do you choose?

?

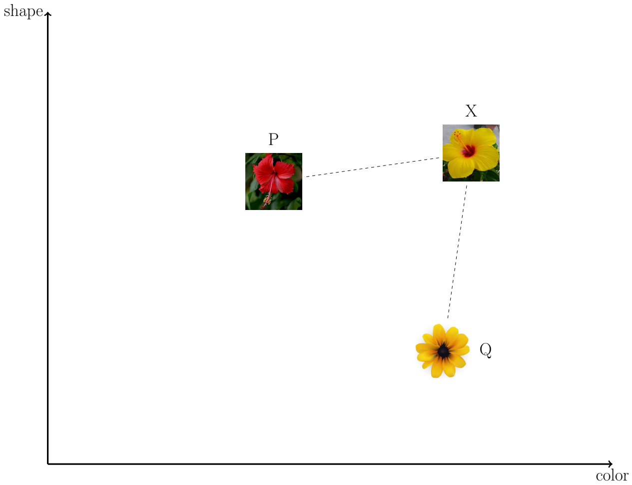
Which category X belongs to?

P and Q are rare classes

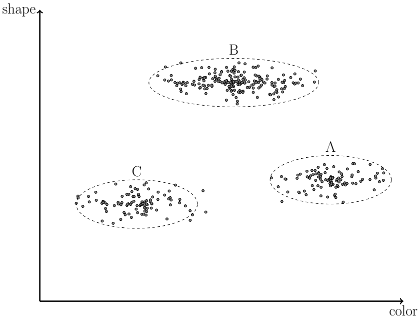


Which category X belongs to?

P and Q are rare classes

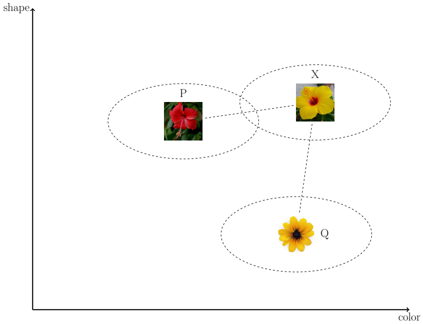


Meta-learning from abundant classes



- Meta-learn that the objects features have more spread in color than shape
- Use more weight for shape than color

Create models for rare classes from meta-model



- This is called Transfer Learning

Shape bias

That is exactly how a child learns to distinguish objects

Quiz

Quiz 03-10

End of Module 03-10

IIT Jodhpur

Biological Vision and Applications
Module 04-01: Feature Integration Theory

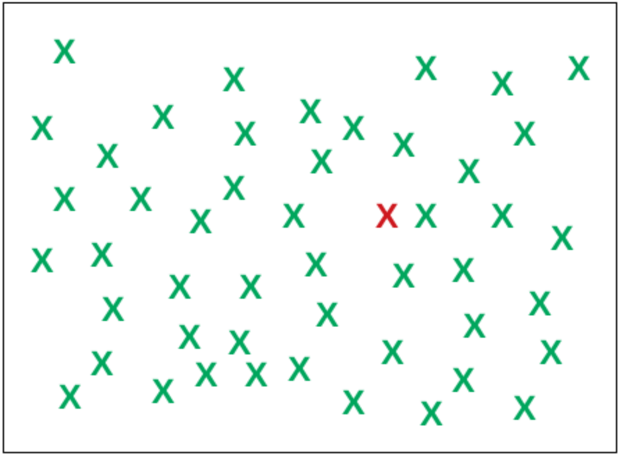
Hiranmay Ghosh

Experiment in visual Search

- You will have to find the red X in the figures in the next two slides
  - There is exactly one red X in each of the figures

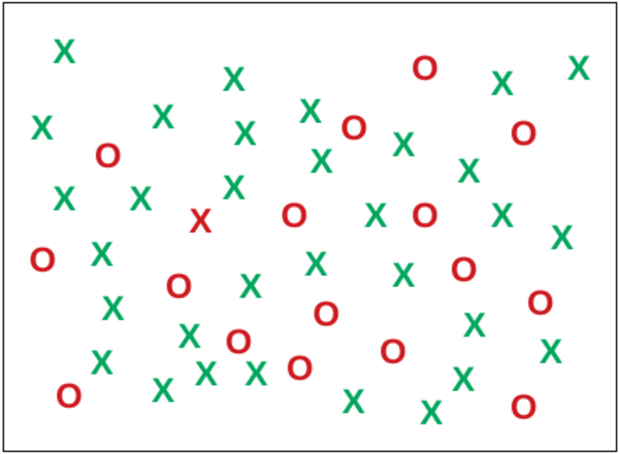
Experiment in visual Search

Find the red X in the figure



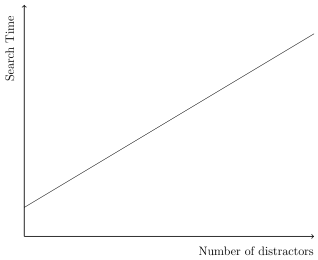
Experiment in visual Search

Find the red X in the figure



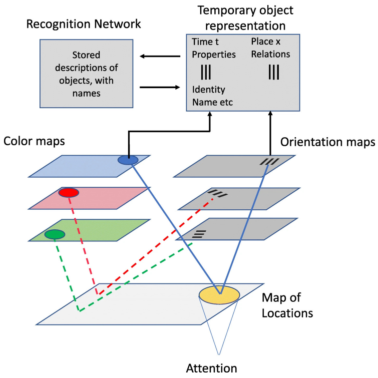
Observations

- When the target is distinguished by color alone, search is almost instantaneous
- When the target is distinguished by color and shape, search takes longer
  - It increases linearly with the number of targets



Triesman's Feature Integration Theory (1980)

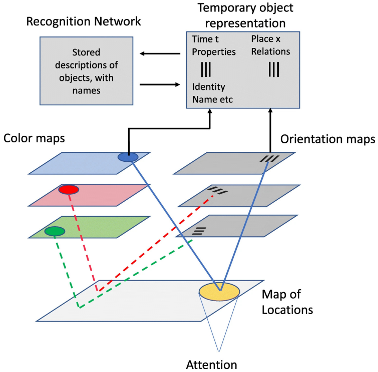
A very simple but elegant theory



- Perceptual process is hierarchical
- **Stage I.** Pre-attentive (early) vision
  - Visual scene encoded on feature dimensions
    - "Automatic"
    - Without any cognitive effort
    - In parallel
  - The locations of objects and mapped
    - "Where" and Not "what"

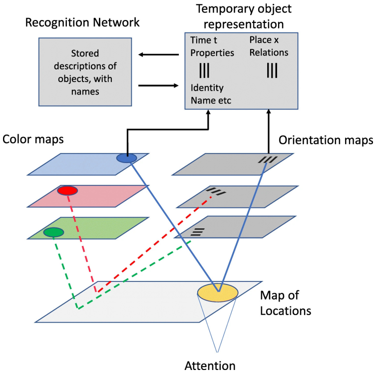
Triesman's Feature Integration Theory (1980)

contd.



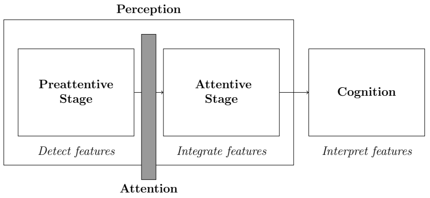
- **Stage II.** Attentive (late) vision
  - Attention "glues" the features together
    - Required for localization
  - Such integrated entities came to be called "visual objects"
    - Conjunction of properties
    - Limited capacitry
    - Features within same attentional focus can be encoded as belonging to the same object

Cognitive process follow perception



- Visual objects compared with descriptions of real objects
  - Objects are detected and localized
- A “scene” is a spatial organization (interaction) of objects
- Events are temporal sequence of scenes (objects and interactions)
  - Within finite temporal bounds (episode)

Vision pipeline



- Are the stages strictly sequential and independent of each other ?
  - Total processing time should be the sum of the individual stages
- Later experiments prove otherwise
  - Relationships between perception, attention and cognition are more complex

Quiz

No quiz for module 04-01

But, we have advance quiz for the next module

End of Module 04-01

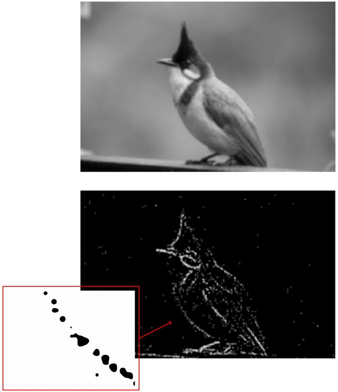
IIT Jodhpur

Biological Vision and Applications
Module 04-02: Perceptual grouping

Hiranmay Ghosh

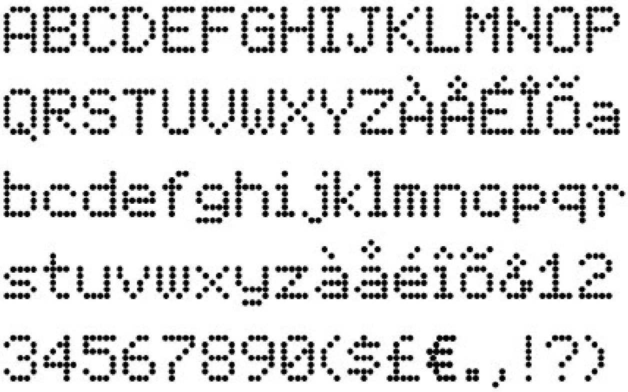
Reconstruction from fragmented contours

- Convolution in eyes result in edge detection
- The process is noisy
  - We do not identify neat object contours
  - Contours are fragmented
  - There are spurious edges
- Human Vision System constructs the object contours through **perceptual grouping**



Seeing the whole, rather than the parts

Dot-matrix printer



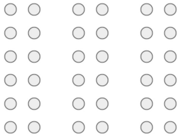
Gestalt psychology

Whole before the parts

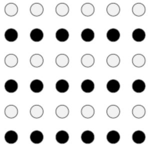
- Do you see the trees first or the forest first ?

Principles of perceptual grouping

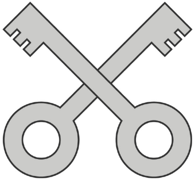
Experiments by Gestalt scientists



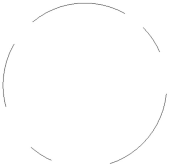
By proximity



By similarity



By continuity



By closure

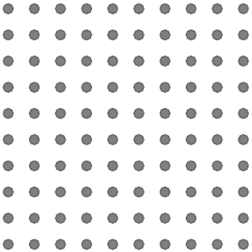
Generic Bayesian formulation [sketch]

for hierarchical perceptual grouping

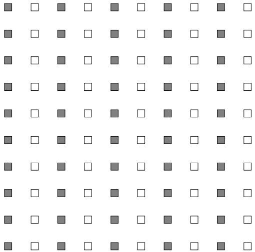
- Data is assumed to be generated by a set of K independent processes c_k
  - Each process represents a “concept” materializing into visible data elements
- Hypothesis space: $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$
  - Each hypothesis is about assignment of data elements to a set of processes
  - Principle of parsimony guides prior probabilities: $P(h_i)$
    - Guided by natural statistics
- Observed data: $d = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
  - Goodness of fit: $P(d \mid h_i)$
- By Bayesian formulation
  - $P(h_i \mid d) = k.P(h_i).P(d \mid h_i)$
  - Choose $h* = \text{argmax}_i P(h_i \mid d)$

What is there is a conflict?

Our experiments



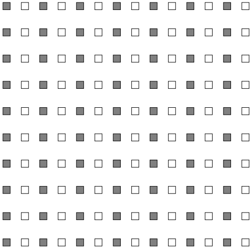
- Equal horizontal and vertical gaps
- 11/21 (52.4%) of you have reported horizontal grouping



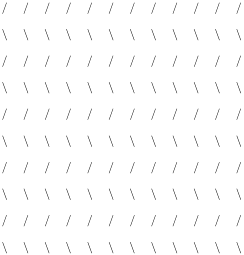
- Equal horizontal and vertical gaps
- 18/21 (85.7%) of you have reported vertical grouping

What is there is a conflict?

Our experiments

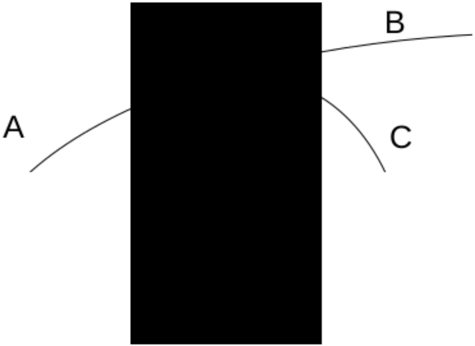


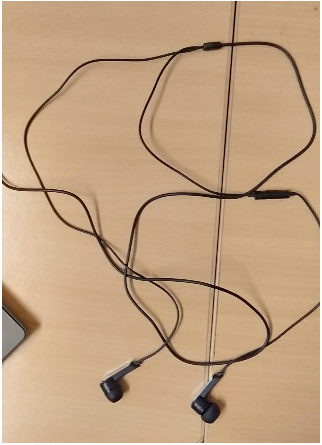
- similarity vs. proximity
  - horizontal gap < vertical gaps
- 15/21 (71.4%) of you have reported verical grouping



- Continuity vs. shape
- 11/21 (52.4%) of you have reported vertical grouping

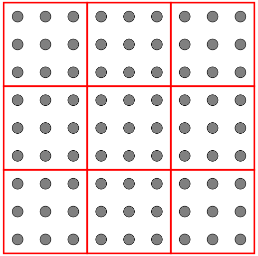
Grouping by continuity



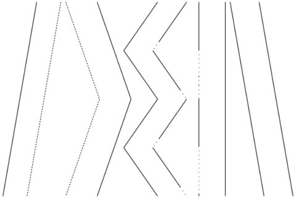


- A $\rightarrow$ B, or A $\rightarrow$ C?

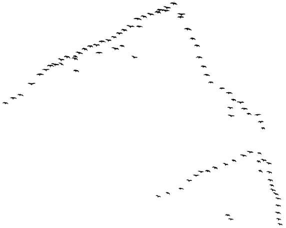
More grouping principles



- Grouping by region
- (x x x)(x x x)



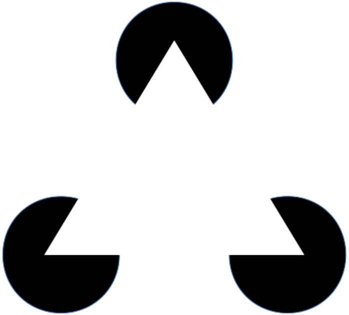
- Grouping by parallelism



- Grouping by common fate (movement)
- Edpuzzle

Closure (completion)

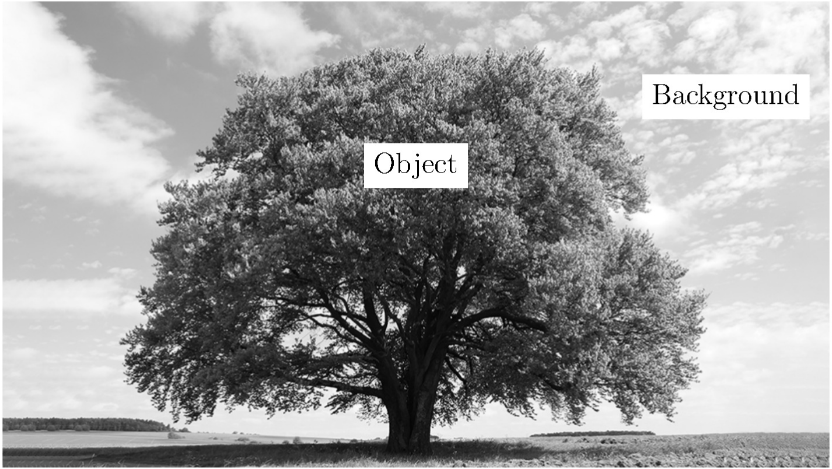
Illusion



- Modal completion
- The white triangle does not exist!
- Kanitza triangle

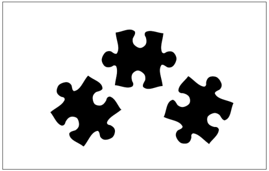
- Amodal completion
- The black triangle is occluded!

Object-ground separation

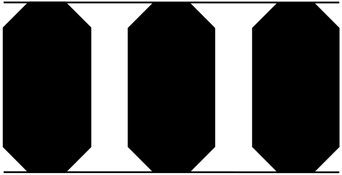


Object-ground separation

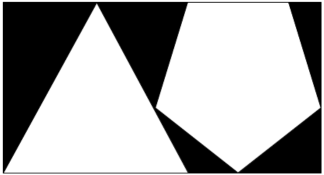
General rules



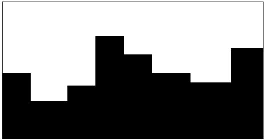
Closed shapes are objects



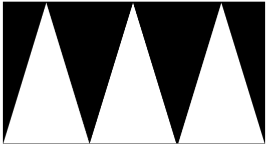
Convex shapes are objects



Symmetric shapes are objects



Shapes at bottom are objects



Shapes with fat bottom are objects

Known shapes are objects

Object-ground separation

Illusion

- Which area is object (foreground) and which area is ground (background)

Bistability

What do you see in the picture ?

See What You See ...

Quiz

Quiz 04-02

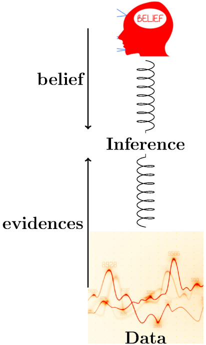
End of Module 04-02

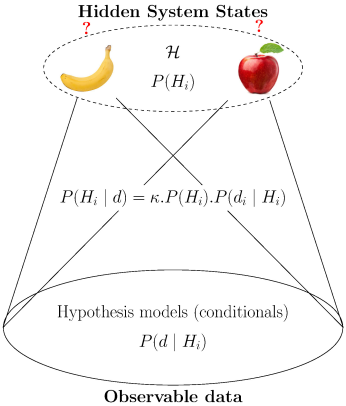
IIT Jodhpur

Biological Vision and Applications
Module 04-03: Object recognition

Hiranmay Ghosh

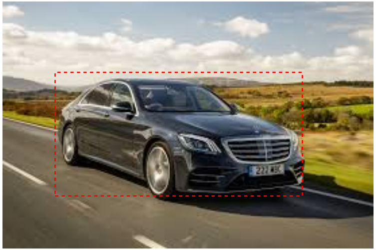
Bayesian Model for object recognition





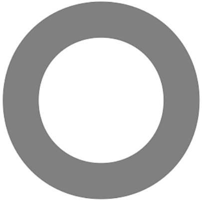
Bayesian Model for object recognition

- $O^* = \text{argmax}_i P(O_i \mid v)$
- when
  - $P(O_i \mid v) = \frac{P(O_i).P(v|O_i)}{P(v)}$
  - O_i = Object hypothesis
  - v = Visual features
- Context contributes to the visual features of the image
  - $v = (v_I, v_c)$ where
    - v_I = Object features
    - v_c = Context features
- In traditional object recognition
  - v_c is minimized
  - $v_I \approx v$



- Can we ignore the context ?

What is the object in this picture ?

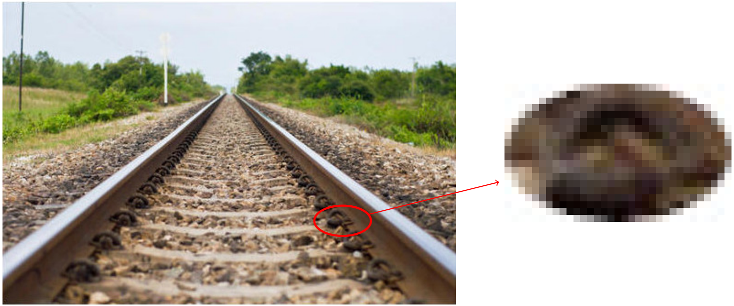


Context matters !

- Seeing the whole provides the cues for identifying the parts

A practical example

Context is especially useful for imperfect images



- Context is especially useful for robust interpretation in imperfect images
  - Ambiguous features, blur, occlusion, clutter, etc.

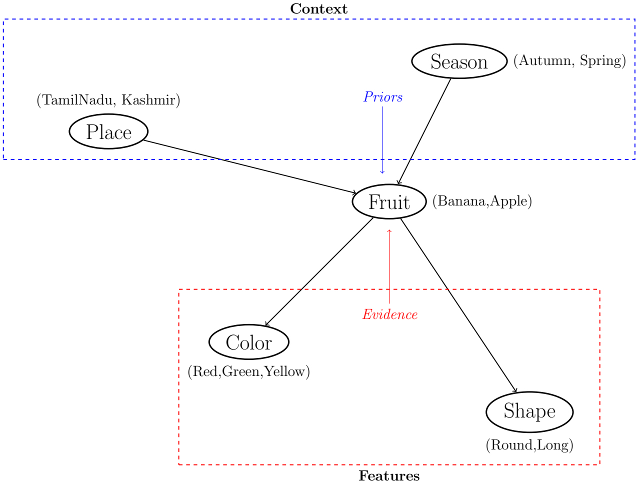
In-context object recognition

We drop the suffix i for convenience

- $P(O \mid v) = k.P(O).P(v \mid O), \qquad [v = (v_I, v_c)]$
- In traditional object recognition $v \approx v_I$
  - $P(O \mid v_I) = k.P(O).P(v_I \mid O)$
- $P(O \mid v_I, v_c) = k'.P(O \mid v_c).P(v_I \mid O, v_c) \qquad$ [*Please deduce*]
  - $P(O \mid v_c)$: Prior probability of the object to appear ... in a specific context
  - $P(v_I \mid O, v_c)$: The model of visual feature of an object ... in a specific context
- We can assume, visual features of an object is independent of context: $P(v_I \mid O, v_c) = P(v_I \mid O)$
  - Has some other significance that we shall analyze in a later lesson

Torralba. Contextual Priming for Object Detection (2003)

Programming assignment 2



The context (in image)

$P(O \mid v_c)$: v_c = visual feature of the context

- $P(O \mid v_I, v_c) = k.P(O \mid v_c).P(v_I \mid O, v_c)$
- Let O not represent just an object class
  - Modeling the visual features with just the class information is too crude
  - Let $O = (o, x, \sigma)$ where
    - o: object class
    - x: location in image
    - σ: appearance (scale, orientation, etc.)
  - $P(O \mid v_c)$ represents an object of a class to appear in a specific location in an image with a certain appearance
- $P(O \mid v_c) = P(o, x, \sigma \mid v_c) = P(\sigma \mid o, x, v_c).P(x \mid o, v_c).P(o \mid v_c)$

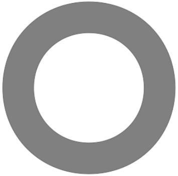
In-context object recognition

Significance of the decomposition

- $P(o \mid v_c)$: Probability of an object class to appear in a context
- $P(x \mid o, v_c)$: Probability of the location where an object class appears in a context
- $P(\sigma \mid o, x, v_c)$: Probability of the appearance of an an object class when it appears in a certain location in an image

- The prior probabilities are determined by the context

How do characterize a context



- Plate is recognized by it's context
- Other objects in the scene creates the context
  - Fork, knife, table-mat
- How do you recognize those objects?
  - A chicken-and-egg problem?

Can we see “forest before the trees” ?

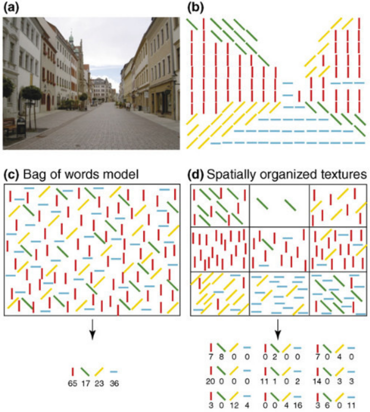
Do the scenes have some distinctive features?





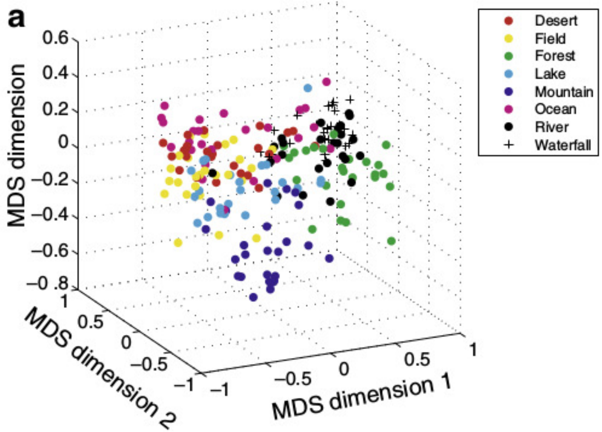

Spatial envelop representation

A holistic representation of a scene layout



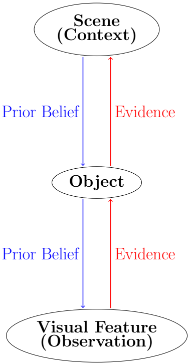
- The edges in a scene constitutes a definite pattern
  - Statistical pattern characterizes a scene
- Recall natural scene statistics
- Happens in early (pre-attentive) vision – fast
- Two types of feature descriptors
  - Global statistics
  - Local statistics
- We skip the detailed mathematical formulation

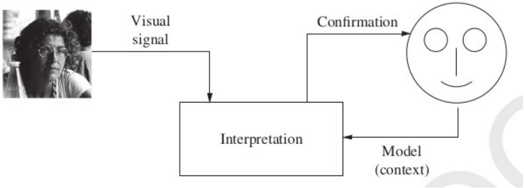
Distinguishing scene classes with spatial envelop representation



Oliva & Torralba. Modeling the Shape of the Scene: ...

Vision as a synthesis of top-down and bottom-up process





- Object recognition is a combination of two processes
  - Top-down: Prior belief (scene context)
  - Bottom-up: Evidence (observation of feature)
- The face model and the face image mutually reinforce belief in each other
- The process is hierarchical

On hypothesis space

Spot the pug

- There are thousands of objects we are familiar with
  - Makes the hypothesis space very large
- Only the hypotheses endorsed by context are analyzed
  - Difficult to detect things at unexpected places

Quiz

Quiz 04-03

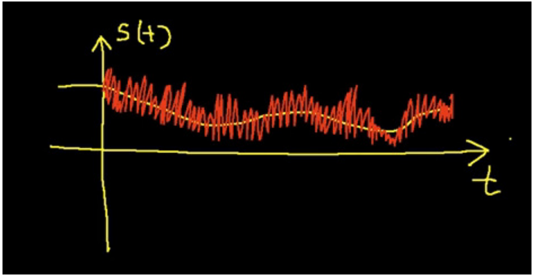
End of Module 04-03

Biological Vision and Applications
Module 04-04: Image quality perception

Hiranmay Ghosh

Signal and Noise

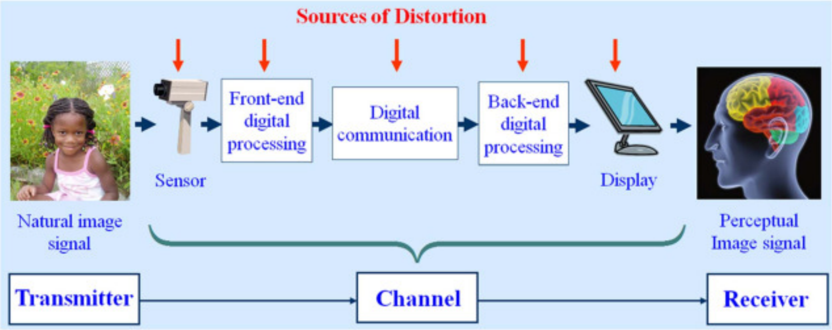
Signal processing view



- Measures of distortion
  - MSE: Mean Square Error (time-domain)
  - SNR: Signal to Noise Ratio (frequency-domain)

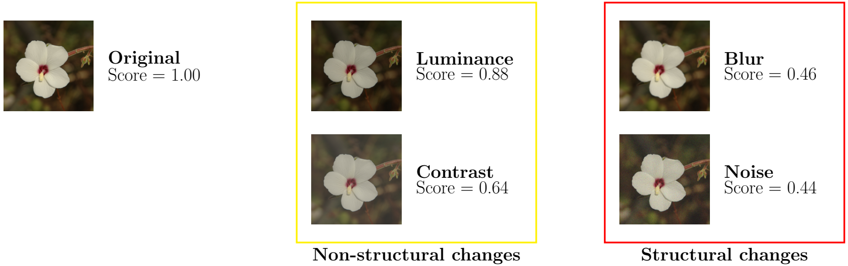
Signal-to-noise ration (YouTube)

How is noise introduced in images



Experimental data

How do we perceive distortion



- Human eye is sensitive to structural information in an image
  - Recall discussions on Natural Scene Statistics
  - Robust against non-structural distortions (normalization mechanism)

Structural Similarity Index Measure

- Measures structural similarity between original and distorted image
- Based on concepts of Natural Scene Statistics
- The original and the distorted images are treated as two probability distributions,
  - Generalized Gaussian Distributions: $p(x)$ and $q(x)$
- $SSIM = KLD(p(x), q(x))$
- A measure of perceptual similarity

- General Gaussian Distribution (Wikipedia) (detailed)
- Kullback-Leibler Divergence

Measurement approaches

- Full reference
  - Both original and distorted images are available
  - Straight-forward
- No reference
  - Original image is not available
    - TV Signal at your home
  - The statistical distribution of the original image is predicted from meta-knowledge about scene statistics
- Reduced reference
  - Some statistical parameters of original image is available
    - Transmitted over a narrow-band channel (assumed to be lossless)
  - The statistical distribution of the original image is predicted by fitting the available parameters to the meta-knowledge

Quiz

No quiz for module 04-04

End of Module 04-04

IIT Jodhpur

Biological Vision and Applications
Module 05-01: Visual attention

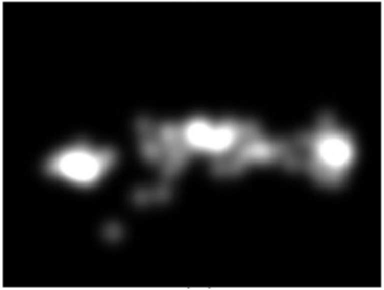
Hiranmay Ghosh

Magnitude of visual data

- Each eye has about 1 million optic nerves coming out of it
- Assume
  - Each nerve carry 1 bit of data and
  - Refreshed every $\frac{1}{10}$th of a second
- Data generated by each eye is 10Mbps (after significant compression)
- Similarly a video camera of modest resolution 1280×768 operating at 30fps generates 700 Mbps
- Now that is a huge data rate
- Do we have to process all that data to understand a scene ?

Attention

We see very little of a scene to understand it's content



- Attention leads to change blindness

Bottom-up Attention

Spontaneously catches attention

- The “red flower” draws spontaneous attention

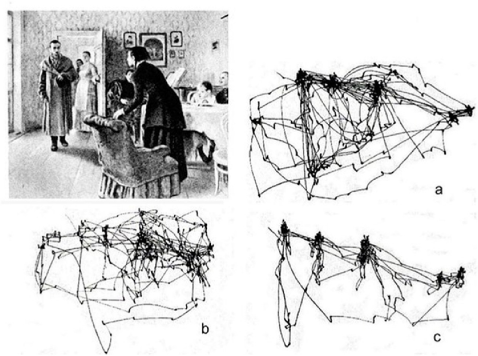
Top-down Attention

Depends on the task / intention of the observer

- Where is my cat? (Visual search)

Yarbus'es experiment

Eye movements depends on the task of the observer



- Attention is dynamic
  - Results in saccades and fixations
- Depends on task, such as
  - (a) Free examination
  - (b) Estimate the material circumstances of the family
  - (c) Give the ages of the people
  - ...

Modeling attention

- Classical approaches
  - Image feature based
  - Location based
  - Object based
- Neural network based approaches
  - We reserve for the future

Quiz

Quiz 05-01

End of Module 05-01

IIT Jodhpur

Biological Vision and Applications Module 05-02: Visual attention: Cognitive model

Hiranmay Ghosh

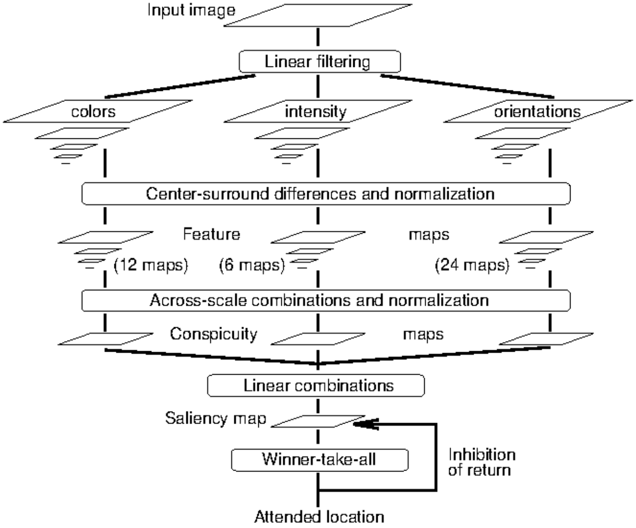
Cognitive Models

Based on Feature Integration Theory

- Motivated by the observations
  - Higher acuity at central vision, lower at peripheral
  - Early vision can distinguish local contrasts
    - Intensity contrast (Dark vs. Bright)
    - Color contrast (Red vs. Green and Blue vs. Yellow)
    - Edge Orientation
  - Features are subsequently integrated
    - Treismann's Feature Integration Theory

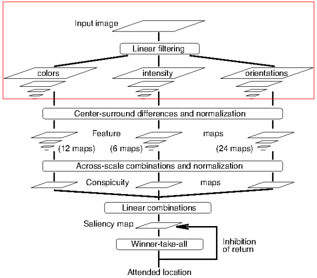
Itti's model (1998)

Overview

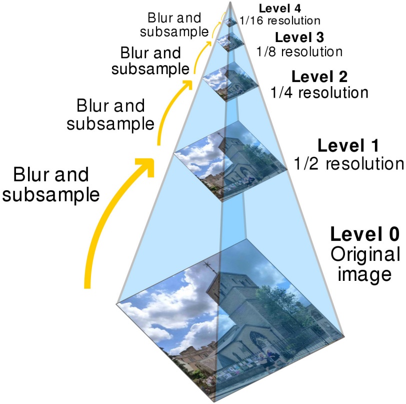


Itti's model: Stage 1

Multi-resolution image analysis

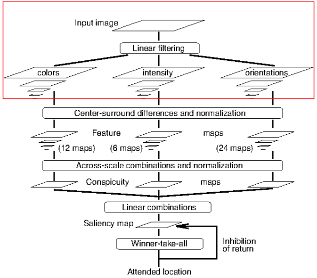


- Multi-resolution analysis of input image
  - Using Gaussian pyramids (9 scales)



Itti's model: Stage 1

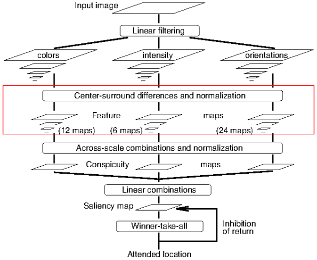
Feature extraction



- For images at each resolution level, 3 features are extracted
  - Color (C): R-G and B-Y contrasts
  - Intensity (I): B-W contrast
  - Edge Orientations (O): 0, 45, 90, 135 degrees
- $2 + 1 + 4 = 7$ features extracted for each resolution level

Itti's model: Stage 2

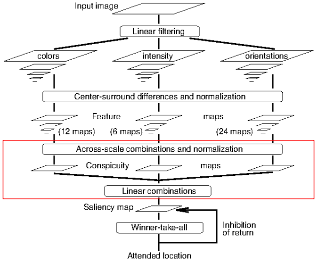
Center-surround operations: Multi-scale feature maps



- Done for each feature
- Center at hi-res, Surround at lo-res
- Scales used:
  - Center: $c = \{2, 3, 4\}$
  - Surround: $s = c + \delta$ $[\delta = \{3, 4\}]$
- Multi-scale differences
  - $\mathcal{F} = \mid F(c) \ominus F(s) \mid$
- 6 scales for each feature
- $7 \times 6 = 54$ "feature maps"
  - Each represents local contrast at a location based on a feature at a certain scale

Itti's model: Stage 3

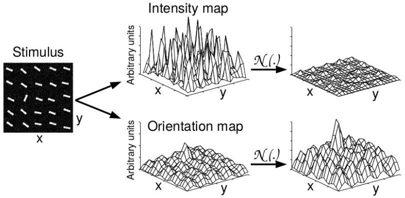
Combining the features: Conspicuity and Saliency Maps



- Feature maps are combined
- Equal weights – normalized $N()$
- Combined in two stages
  - Intra-feature-class, giving three *conspicuity maps*
    - $\bar{I} = \bigoplus_{c,s} N(I(c,s))$
    - $\bar{C} = \sum_{RG,BY} \bigoplus_{c,s} N(C(c,s))$
    - $\bar{O} = \sum_{\theta} \bigoplus_{c,s} O(c,s)$
  - Inter-feature-class, giving the final *saliency map*
    - $S = \bar{I} + \bar{C} + \bar{O}$

Itti's model: Stage 3

Normalization

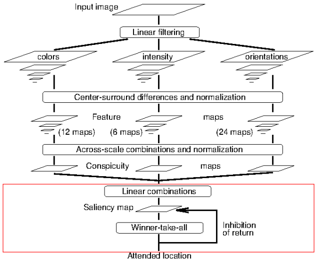


Maxima	**6.00**	**7.00**	**5.00**	**6.00**	**5.00**
Normalized	**0.04**	**0.05**	**0.03**	**0.04**	**0.03**
Maxima	**6.00**	**20.00**	**5.00**	**6.00**	**5.00**
Normalized	**0.16**	**0.53**	**0.13**	**0.16**	**0.13**

- Two reasons to normalize
  - Features are at arbitrary scale
  - Normalize to a fixed range $[0 \dots M]$
- Some feature may have many nearly equal peaks, indicating texture
  - Find the global maximum M
  - Compute the average of all other local maxima $\bar{m}$
  - Multiplying the map by $(M - \bar{m})^2$

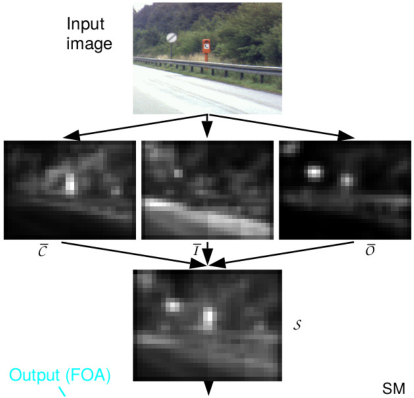
Itti's model: Stage 4

"Winner take it all" and "Return Inhibition" policies

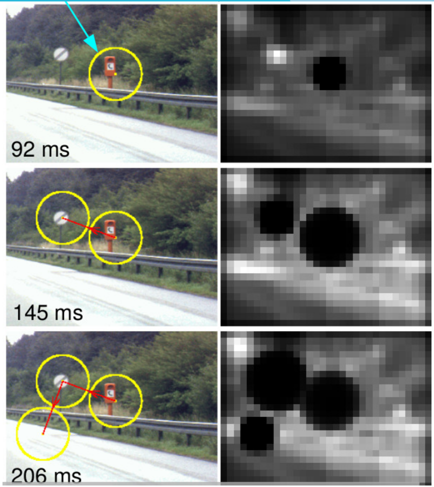


- Winner-take-it-all policy
  - The image location with highest saliency attracts attention
  - All other locations are ignored
- Return Inhibition policy
  - Attention never returns to a location once attended
  - The neurons at the attended place tires out.
  - Attention moves to the location with next highest salience.

Results



Input image
$\overline{C}$
$\overline{I}$
$\overline{O}$
S
Output (FOA)
SM



92 ms
145 ms
206 ms

Quiz

Quiz 05-02

End of Module 05-02

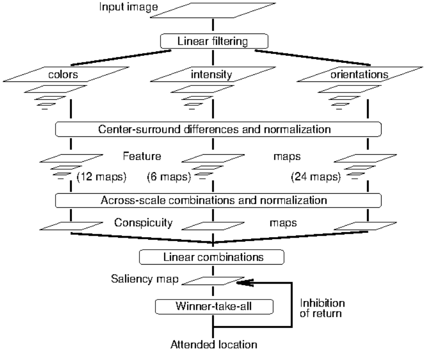
IIT Jodhpur

Biological Vision and Applications
Module 05-03: Visual attention: Extensions to Itti's model

Hiranmay Ghosh

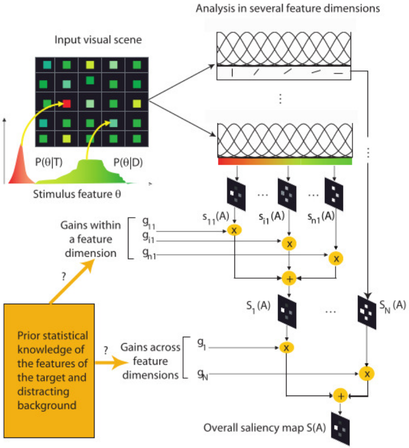
Itti's model

Recap



- Based on cognitive theories of early vision
- Features used: Color, Intensity and Orientations
  - Equal weights to all features
- Models bottom-up attention
- Provides static saliency map
- Eye movement guided by
  - Winner Take All policy
  - Return Inhibition policy
- Remains a reference model till date
- WTA and RI policies are common to all classical models

Adaptation to top-down attention

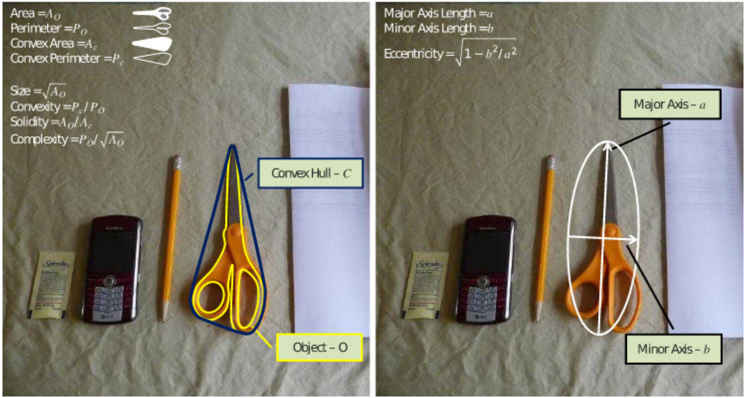


- Visual search task
- Weights assigned to features based on task requirement
- Weights learned from statistical features of target and distractors
- Inflexible

Extension of feature set

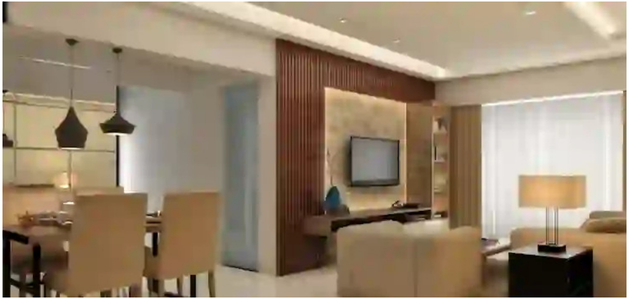
Object level attributes

- Recall what is likely to be a foreground object
  - Local motion (for video)
  - convex-ness ...



Extension to feature set (contd.)

What draws human attention? – Rethinking the principles



Semantic features

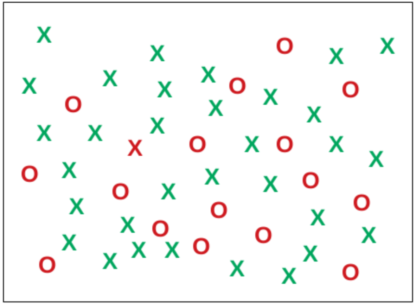
- Semantic features
  - Human face and emotions
  - Text
  - Man-made objects designed to be watched (TV, clock, ...)
  - Objects with sound, smell, taste, touch attributes
  - Objects interacted with (touched or gazed upon by) humans (a computer mouse, ...)
  - ...

Early fusion vs. late fusion

When to fuse the conspicuity maps?

- Early fusion
  - As in Itti's model
  - Fused immediately after normalization
  - Overall saliency map created after fusion
- Late fusion
  - Create saliency map based on one feature
  - Fuse conspicuity maps from the other features for the competing locations
    - One at a time
  - Computationally more efficient

More on Late Fusion



- Example: In finding red X problem,
  - Based on color saliency, get all red objects first
  - Fuse shape based saliency next
    - Only where we find red
- What is the priority sequence of features
  - General agreement on “color first”
  - No clear consensus on others

Quiz

No quiz for module 05-03
But, we have an experiment

End of Module 05-03

Biological Vision and Applications
Module 05-04: Probabilistic models

Hiranmay Ghosh

Information theoretic model

- Challenges à-prior defined features for attention in Cognitive models
- Based on Shannon's information theory:
  - The event that is least likely to occur has the maximum information value
  - Self-information of an event x: $-\log P(x)$
- Image region that is least likely to occur in an image is the most salient one
- How to decide what is least likely to occur?

Shannon's Information theory

- An image region (signal) is a manifestation of some underlying hidden processes
- It is a weighted sum of contributions from each process
  - The processes are independent of each other
  - The processes are hidden
  - The weights are unknown
- We resort to Natural Scene Statistics
  - Observe a large number of images
  - Learn the processes and the weights
- Given a new image, we would expect each region to be a weighted sum of contributions from some of these features

What is unexpected ?

- Model a new image with the learned features
  - Select features for best fit over all image regions
- There will be some outlier regions, which do not fit
  - Has least probability to occur
  - Has most information value
- These image regions are the salient ones
  - Lower is the probability, the higher is the saliency
- Bruce. Features that draw visual attention: ...
- Bruce & Tsotsos. Saliency Based on Information Maximization (2006)

Bayesian model

Based on "surprise" – brings in experential factor

- Image regions (observed data) are caused by some hidden states (features)
- Bayesian definition of surprise
  - M: a continuous range of states (features that manifests as image regions)
  - $p(m)$: the prior pdf for states (characterizes alternate models)
  - $p(m \mid D)$: the posterior pdf for the states, after experiencing some data D
- The surprise factor of the data D is defined as
  - The difference between the posterior and the prior pdfs for M
    - $S(D) = KLD(p(m), p(m \mid D)) = \int_m p(m).\log\frac{p(m)}{p(m|D)}.dm$

Surprise

... contd.

- The surprise factor of the data D: $S(D) = \int_m p(m).\log\frac{p(m)}{p(m|D)}.dm$
- Baye's Theorem: $p(m \mid D) = \frac{p(m)P(D|m)}{P(D)}$
- Using Baye's Theorem and simplifying $[\int_m p(m) = 1]$
  - $S(D) = \log P(D) - \int_m \log P(D \mid m).dm$
- The model of the environment builds incrementally as the agent observes
  - For repeated observations, posterior of one observation is the prior for the next
  - Leads to change awareness

Itti & Baldi. Bayesian Surprise Attracts Human Attention (2005)

Outlier and surprise

Criticism to information theoretic model

- Outlier D: $P(D \mid M_{\text{best}})$ to have low value
- Suppose $P(D \mid M')$ is low for all possible alternate models M'
- D has little informative value for discriminating the models for best explanation of the observation
- Cannot be treated as "surprise" and cannot be considered to be a guiding factor for attention

Eye movement ?

- Fixations and saccades are guided by WTA and RI policies as in Cognitive models

Quiz 05-04

End of Module 05-04

IIT Jodhpur

Biological Vision and Applications
Module 05-05: Context-based model

Hiranmay Ghosh

Context-based Model

A comprehensive model for top-down + bottom-up attention

- We have seen the equation earlier
  - $P(O \mid v_l, v_c) = \frac{1}{P(v_l \mid v_c)}.P(v_l \mid O, v_c).P(O \mid v_c)$
- $P(v_l \mid v_c)$:
  - Independent of object hypothesis O
  - Represents probability of the observed feature v_l in context v_c
  - It's inverse represents bottom-up saliency
    - Information-theoretic / Bayesian models ... bring in experiential factor

Top-down saliency

- We shall now contrate on the numerator:
  - $N = P(v_l \mid O, v_c).P(O \mid v_c)$
- $O = (o, x, \sigma)$ (class, location, appearance)
  - $P(O \mid v_c) = P(\sigma \mid x, o, v_c).P(x \mid o, v_c).P(o \mid v_c)$
- $N = P(v_l \mid O, v_c).P(\sigma \mid x, o, v_c).\underline{P(x \mid o, v_c).P(o \mid v_c)}$
  - $P(o \mid v_c)$: Prob of an object class to appear in a context
  - $P(x \mid o, v_c)$: Prob of an object class to appear at a certain location in a context
    - ... given that it appears
- $P(x \mid o, v_c).P(o \mid v_c) = P(o, x \mid v_c)$:
  - Represents the task-specific context-driven saliency of <u>location</u>

Appearance model

- We are left with the terms: $P(v_l \mid O, v_c).P(\sigma \mid x, o, v_c)$
  - $P(\sigma \mid x, o, v_c)$: Scale and appearance of the object at a certain location
    - ... given that it appears at a certain location
  - $P(v_l \mid O, v_c) = P(v_l \mid O) = P(v_l \mid o, x, \sigma)$: The expected visual features
    - ... given that the object appears at a certain location with a certain appearance

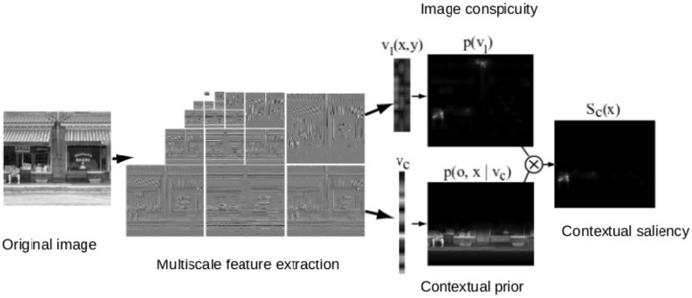
In Summary

- $P(O \mid v) = P(O \mid v_l, v_c) = \frac{1}{P(v_l|v_c)}.P(v_l \mid O, v_c).P(O \mid v_c)$
- Substituting (o, x, σ) for O
  - $\mathrm{P(O} \mid v) = \underline{\frac{1}{P(v_l|v_c)}}. \underline{P(v_l \mid O, v_c).P(\sigma \mid x, o, v_c)}. \underline{P(x \mid o, v_c).P(o \mid v_c)}$

- Bottom-up saliency (s_b): $\frac{1}{P(v_l|v_c)}$
- Top-down saliency (s_t): $P(o, x \mid v_c) = P(x \mid o, v_c).P(o \mid v_c)$
- Appearance model (a): $P(\sigma \mid x, o, v_c).P(v_l \mid o, x, \sigma)$

- Overall saliency (where): $s_b \times s_t$
- Feature to look for (what): a

Local Computations



Global Computations

Quiz

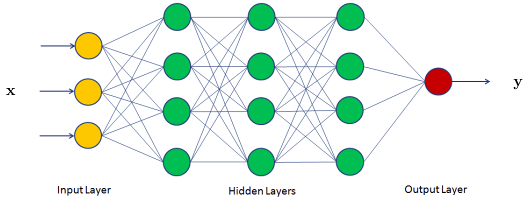
Quiz 05-05

End of Module 05-05

Biological Vision and Applications
Module 05-07: Introduction to neural networks

Hiranmay Ghosh

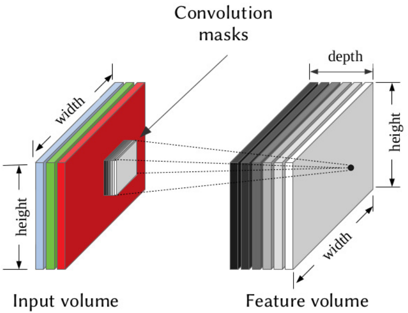
Neural Networks



- Feed-forward network – back-propagation algorithm for training
- Transfer function: $y = W.x$
- W is a constant: deterministic output
  - no learning from "experience" in deployment stage
- For a 640×480 color image
  - Number of input nodes = 927,360
  - Large number of parameters to be learned

Convolutional Neural Networks

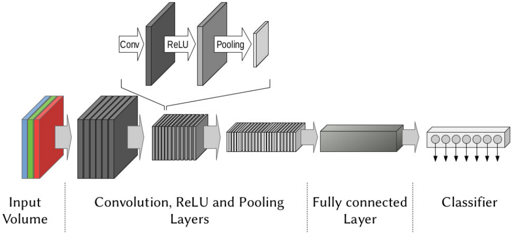
Why convolution?



- 2D organization exploits
  - Spatial context of a location in 2D
  - Identical operations repeated over the different spatial regions
- Drastic reduction in model parameters
  - For a 3×3 convolution filter, only 27 parameters to learn
  - Independent of image size

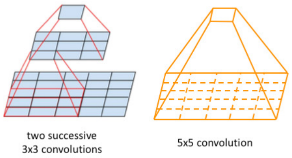
Convolutional Neural Networks

Structure

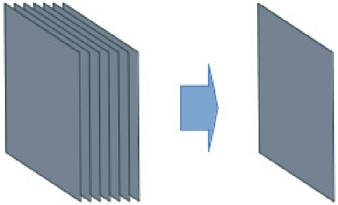


- Same operation to be repeated over different receptive fields
  - Do not need so many parameters
- Architecture motivated by early vision
  - Convolution: Aggregates information from receptive field
  - Filtering (ReLU): Non-linear transformation
  - Pooling (max): Reduces information volume

On filter sizes

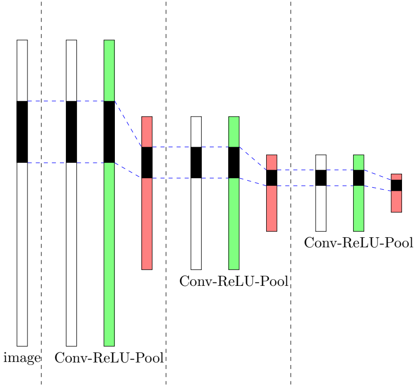


- A bank of two 3×3 filters in succession has a receptive field of 5×5
  - Can implement identical transfer function
- Which one would you prefer?



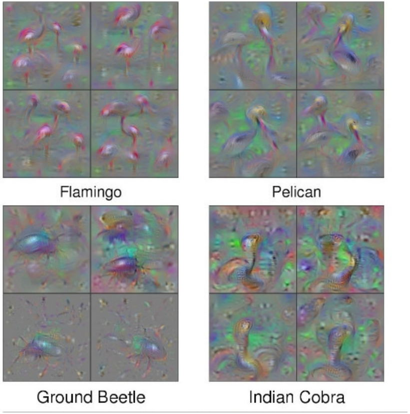
- Filter size $= 1$
- For "flattening" the layers
  - $y(i,j) = \sum_k w_k . x_k(i,j)$

Progressive abstraction



- Each location at any layer of a CNN holds information about some locality of the image
- A location in a deeper layer covers more visual field of the image than a shallower layer
  - A deeper layer incorporates more context than a shallower layer
- Visual information is progressively abstracted
- Depth of layer increases with the depth of the network

Does CNN really do progressive abstraction ?



Yosinski, et al. Understanding neural networks through deep visualization (2015)

Some notable CNN implementations (2012 – 2015)

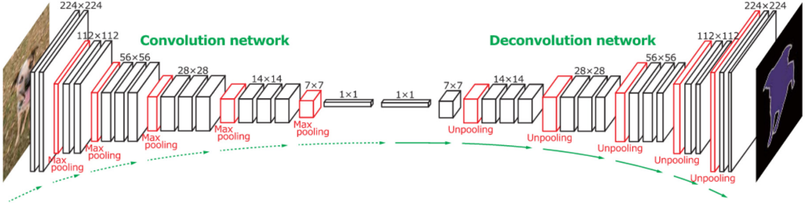
These implementations are reused in different contexts

- AlexNet
- VGG
- ResNet
- GoogleNet

Architecture comparisons

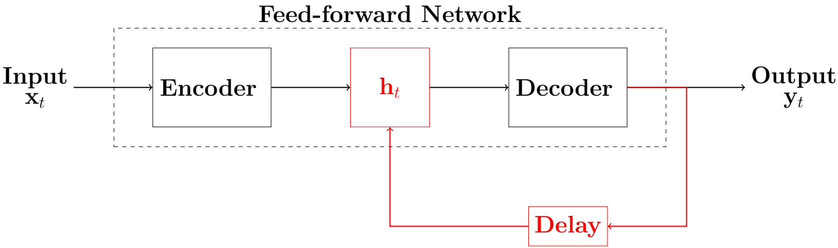
Fully Convolutional Network (FCNN)

Used for Image Segmentation



Deconvnet

Recurrent Neural Network (RNN)



- RNN incorporates a feedback loop (with delay)
- Transfer function
  - $h_t = f(W_1.x_t + W_2.y_{t-1})$
  - $y_t = g(W_3.h_t)$
  - h accumulates experience
- Tool for sequence processing tasks

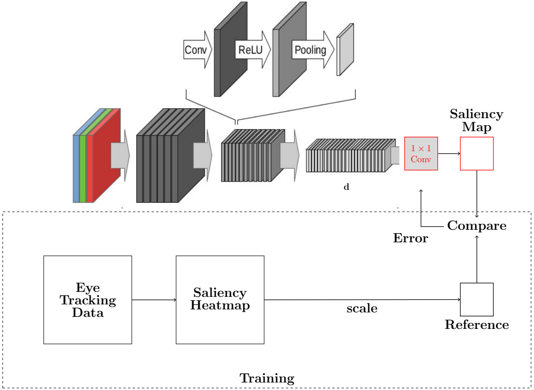
Quiz 05-07

End of Module 05-07

Biological Vision and Applications
Module 05-08: NN based attention models

Hiranmay Ghosh

Basic Architecture



- Bottom-up or top-down attention?

Attention and object detection

- Use CNN pre-trained for object detection
  - Not enough training data for saliency
  - Objects lead to saliency
- In neural network based architectures
  - Attention and object detection complement each other
  - Find salient locations (where objects are likely to be there)
  - Detect objects at those locations

Soft attention vs. hard attention

- Soft attention
  - Graded saliency values for different image locations
  - Fixation traverses from location with highest saliency to lowest
- Hard attention
  - Binary saliency values
  - Fixation at the region with saliency
- NN based attention models generally use hard attention
  - One or very few “salient” objects in a scene
  - A binary classifier (SVM / Softmax) is added at the end

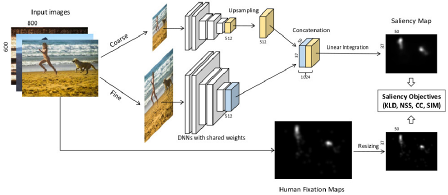
Objects are salient

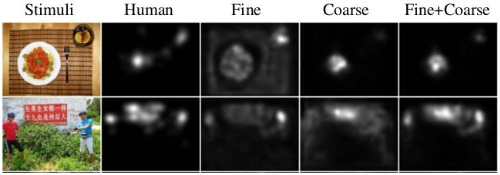
Saliency of nearby regions should be close to each other

- The approach discussed computes location based saliency
- The image can be divided into 'superpixels'
  - Areas of near uniform color/texture
- Adjust saliency values to encourage locations in nearby superpixels to have homogeneous saliency
  - Something like graph-cut algorithm
  - Minimize $\sum_i (s_i^{new} - s_i)^2 + \sum_{i,j} w_{ij}(s_i^{new} - s_j^{new})^2$
  - Weights w_{ij} depend on physical distance
    - Optimal weights are learned

Multi-scale analysis

SALICON: Saliency in Context





- Coarse level captures context; fine level captures local contrasts
- Usually 2 or 3 levels of resolution is found to be sufficient

Quiz 05-08

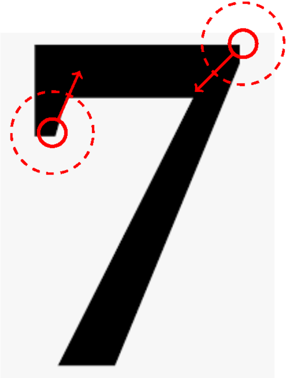
End of Module 05-08

IIT Jodhpur

Biological Vision and Applications
Module 05-09: Recurrent attention models

Hiranmay Ghosh

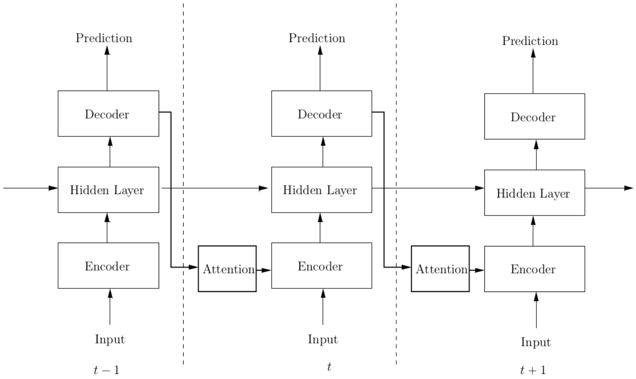
Saliency is dynamically constructed



- We look at a small part of a scene at a time
- Where we look at next depends on what we see
  - … plus, the task at hand
- Saliency map of a scene is not computed in one go
  - Constructed dynamically over time
    - As and when needed … Just in time
  - Saliency map for the whole image is never built
- Peripheral vision guides the direction of eye movement

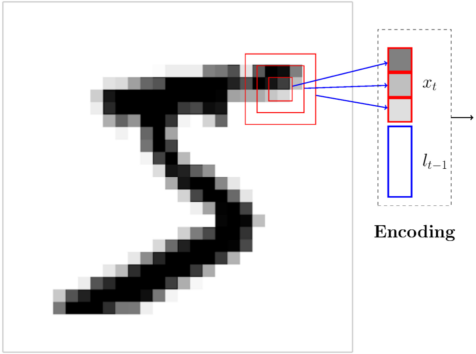
EdPuzzle assignment

Attention-based RNN Architecture



- RNN and the “Attention” module are trained together

Implementation example

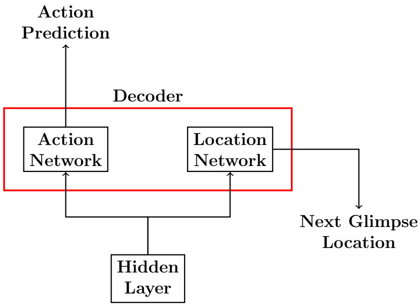


- Encoding
  - Glimpse: Encoded representation of visual field
  - Glimpse Network:
    - Image data + Location (x_t, l_{t-1})
    - Encoded to some internal representation with an NN

- Where do you look at the first glimpse? $l_0 = ?$

Mnih, et al. Recurrent models of visual attention (2014)

Decoder



- Each of Action and Location Networks is an NN
- "Action" can be different in different contexts:
  - Predicting the object
    - number, in our example
  - Driving a car
  - ...

Training

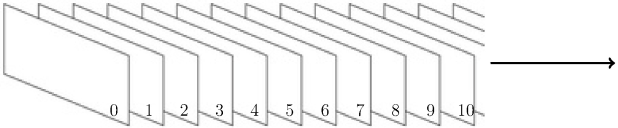
Reinforcement Learning

- Model for "hard attention"
- Training for optimal saccades
  - Training based on back-propagation does not work
  - Reinforcement learning used
  - Reward after each time-step
  - Biological system might follow similar "reward" based learning mechanism
- In the case of object recognition
  - Reward $r_t = 1$ if the object is classified correctly at time step t
  - $r_t = 0$ otherwise
- Positive reward is sparse
- System tries to maximize $\sum_t r_t$ over time

Discussions

- Attention and object recognition complements each other
- Example of “life-long learning”
- Network trained on a few patterns performs well for other patterns with little training
  - Example of transfer learning
- Robust against distractors (noisy patches on the image)

Recurrent Attention for Video

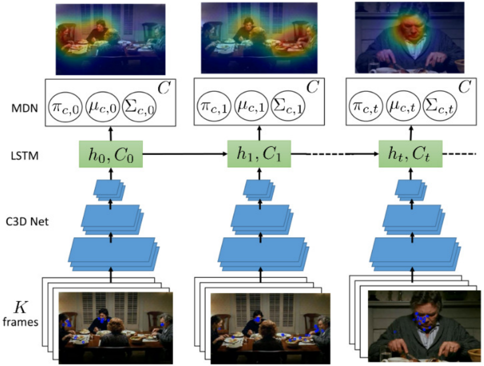


Why processing video frame by frame does not work ?

- Motion information is lost
- Saliency map for each frame depends on the earlier frames
- Too much data to be processed
  - There are lots of redundancies in video data (over successive frames)

Recurrent Attention Model for Video

Recurrent Mixture Density Network



- Soft attention model is used
- Prediction in the form of a GMM over space
  - There can be multiple salient objects

Bazzani & Larochelle. Recurrent mixture density network ... (2017)

https://www.youtube.com/watch?v=aX0wc17nx_s

Discussions

- Wasteful processing
  - Same frame processed multiple times
  - Alternate approach uses two layers of LSTM
    - Lower layer: short-term temporal variations (motion features)
    - Upper layer: long-term history learns to predict saliency
- Camera motion vs. object motion
  - Object motion matters
    - FG–BG separation
    - Assign weights to FG

Quiz

Quiz 05-09

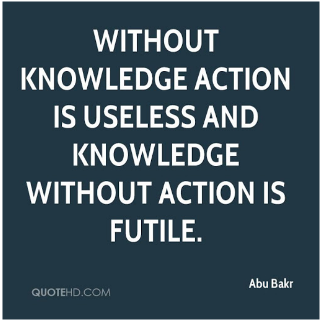
End of Module 05-09

IIT Jodhpur

Biological Vision and Applications
Module 06-01: Cognitive systems

Hiranmay Ghosh

Goal of Cognitive Systems



- We want to build systems that can intelligently act

Cognitive Systems

- Examples
  - Autonomous cars
  - Social robots

- Need to have intellectual capabilities comparable to humans
- The systems must interact autonomously with the environment
  - Must have a goal to fulfill
  - Must work without human intervention
  - Needs to understand the environment and act
  - Combine cognition with action
  - Need to be versatile: work in multiple environments

Autonomous Agents (Intelligent Agents)

Properties

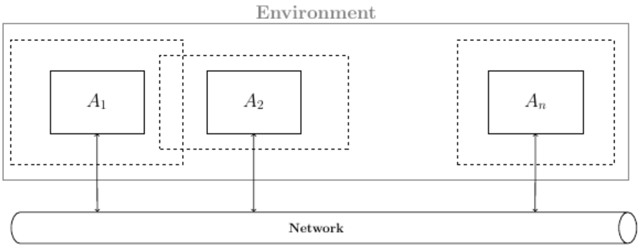
- Autonomy: Can act independently without human intervention
- Goal directed: Has a goal to achieve – tries to maximize some benefit
- Interactive: Can sense and act on the environment
- Social capability: Can communicate with humans and other intelligent agents
- Knowledge: Maintains a model of the world (environment + itself)
- Learning: Stores experience and improve it's performance with experience

- A cognitive system comprises one or more interacting autonomous agents

An agent is "situated" in an environment

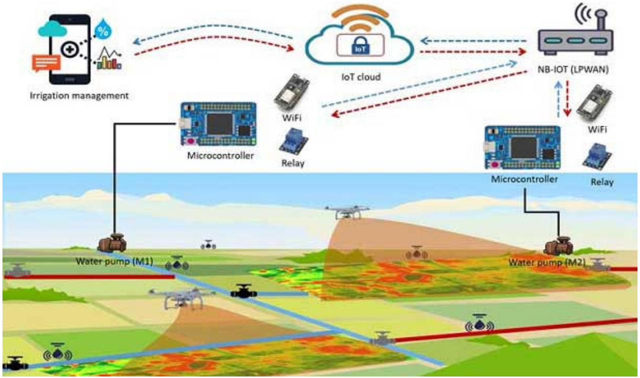
- Physical / embodied: built on dedicated hardware
  - Autonomous car, drones, ...
  - Robotic vacuum cleaner
  - Industrial and Social robots
  - A/C thermostat
- Software / non-embodied: hosted on commodity hardware (laptop/mobile)
  - E-Mail spam filter
  - AI based chess player
  - Recommendation engines

Multi-Agent Systems



- All the agents inhabit the same world
  - Their (local) environments may be different
- Agents communicate with each other

IoT based system as an example of Multi-Agent System



- Smart agriculture

Quiz

No quiz for module 06-01

End of Module 06-01

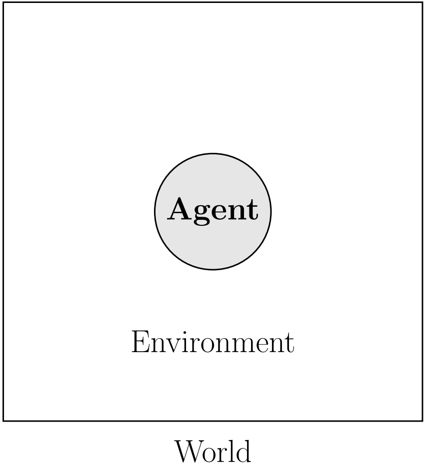
IIT Jodhpur

Biological Vision and Applications
Module 06-02: Agents and Environment

Hiranmay Ghosh

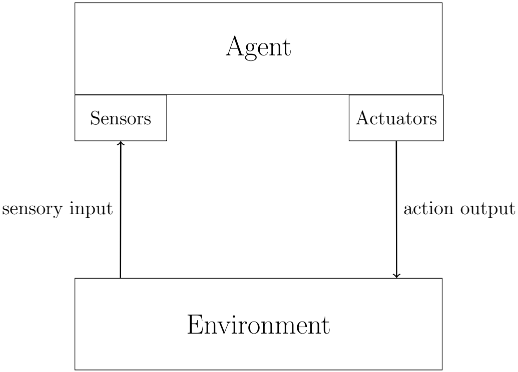
A situated agent

Agent, Environment and the World



An agent interacts with it's environment

Sensing and effecting (or, actuating)



Characterizing the environment

- Accessible vs. Inaccessible
  - Accessible: Agent has complete information about the environment
  - Inaccessible: Agent has partial information about the environment
- Deterministic vs. non-deterministic
  - Deterministic: An action by an agent has a guaranteed effect on the environment
  - Non-deterministic: The change in the environment cannot be guaranteed
- Episodic vs. non-episodic
  - Episodic: Distinct episodes – no link between episodes
  - Non-episodic: No such boundary – all events in an agent's life are linked

Characterizing the environment

Continued

- Static vs. Dynamic
  - Static: Does not change except for that caused by the agent's action
  - Dynamic: Changes with time, even if an agent does not act
  - Semi-static: Can change by action of some other agent
- Discrete vs. Continuous
  - Discrete: The environment states are discrete
  - Continuous: The environment states are continuous

Modeling of Agent and Environment

- An agent can execute a finite set of actions
  - Let the action repertoire of an agent be $A = \{a_1, a_2, \ldots, a_m\}$
- The environment has a finite number of states
  - Let the states of an environment be $S = \{s^1, s^2, \ldots, s^n\}$
- At any given point of time t, the environment is characterized by a sequence of states that it has traversed
  - $S^* = (s_0, s_1, \ldots, s_t)$ where
    1. $s_i (i = 1, \ldots, t) \in S$
    2. s_0 represents the initial state of the environment

Modeling of Agent and Environment

Agent

- In general, an agent translates a sequence of environment states to an action
  - $action(s_0, s_1, \ldots, s_t) \rightarrow a$
- Behavioral model of an agent is defined by the action function
  - $action : S^* \rightarrow A$, where
    - S^* denotes a sequence of environment states
- As a special case, the action of an agent may be determined by the current state alone
  - $action : S \rightarrow A$

Modeling of Agent and Environment

Environment

- An action changes the state of an environment
  - There can be uncertainty in the state that the environment will transit to
- The change in environmental state can be modeled as
  - $env(s, a) \rightarrow \rho(S)$, where
    - $\rho(S)$ denotes an set of states
    - Possible outcomes of the action
    - Often attributed by a probability distribution
- If $\forall s, a : env(s, a)$ is a singleton
  - The environment is deterministic
  - ... Otherwise, it is non-deterministic
- Behavioral model of environment
  - $env : S \times A \rightarrow \rho(S)$

Modeling of Agent and Environment

History and Characteristic Behavior

- A history of a system $(\mathcal{A}, \mathcal{E})$ is defined as the sequence of environment states and agent actions
  - $h(\mathcal{A}, \mathcal{E}) = s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \ldots s_i \xrightarrow{a_i} s_{i+1} \ldots$, when
    1. $\forall i : a_i = action(s_0, s_1, \ldots, s_i)$
    2. $s_{i+1} \in env(s_i, a_i)$

- Characteristic behavior of an agent $\mathcal{A}$ in environment $\mathcal{E}$
  - $\mathcal{H}(\mathcal{A}, \mathcal{E}) = \{h(\mathcal{A}, \mathcal{E})\}$
  - ... (set of all possible histories)

Modeling of Agent and Environment

Summary

- Agent: $\mathcal{A} \equiv action : S^* \rightarrow A$
- Environment: $\mathcal{E} \equiv env : S \times A \rightarrow \rho(S)$
- History: $h(\mathcal{A}, \mathcal{E}) \equiv s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \ldots s_i \xrightarrow{a_i} s_{i+1} \ldots$
- Characteristic behavior of $\mathcal{A}$ in $\mathcal{E}$: $\mathcal{H}(\mathcal{A}, \mathcal{E})$
  - where, $\mathcal{H}(\mathcal{A}, \mathcal{E}) = \{h(\mathcal{A}, \mathcal{E})\}$
- Two agents $\mathcal{A}_1$ and $\mathcal{A}_2$ are said to be behaviorally identical with respect to environment $\mathcal{E}$, iff
  - $\mathcal{H}(\mathcal{A}_1, \mathcal{E}) = \mathcal{H}(\mathcal{A}_2, \mathcal{E})$

Modeling of Agent and Environment

Invariant property

- A property P is called an invariant property of the characteristic behavior $\mathcal{H}(\mathcal{A}, E)$, iff
  - P holds good for all $h(\mathcal{A}, \mathcal{E}) \in \mathcal{H}(\mathcal{A}, \mathcal{E})$
- Example: a car does not collide with another while driving
- An invariant property in one environment does not necessarily hold good in another environment
  - A car drives on the left side on Indian roads

Abstract architecture and Concrete architecture

- Abstract architecture
  - Deals with internal states, state-transitions of an agent
  - Does not worry about how it is implemented
- Concrete architecture
  - Deals with representations and implementations
  - Logic, Bayesian reasoning, Petrinet, ...

Quiz

Quiz 06-02

End of Module 06-02

Biological Vision and Applications
Module 06-03: Reactive Agents

Hiranmay Ghosh

Reactive Agent

- An agent senses the environment and reacts to it
  - An autonomous car applies the brakes on seeing a pedestrian in the front
- Reactive behavior is spontaneous and immediate
  - Pre-attentive, without deliberation

Can we model a common room AC as a Reactive Agent?

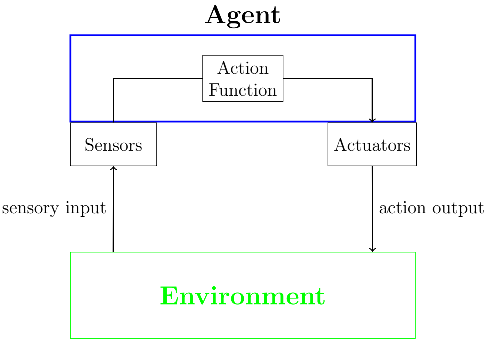
- Properties of a room AC
  - Autonomous: works without human intervention
  - Interactive: Senses and acts on the environment
  - Goal: to keep the room cool (implicit)
  - Social capability: communicates with humans
  - Knowledge: Knows how to do it's job (implicit)
  - Learning: Not applicable
- By definition, it is an agent
- It may an overkill to model such simple system as an agent
- Nevertheless, we shall use it as an example in this class

Characterizing the environment

- Accessible? ... **Yes**
  - We assume that the temperature sensor provides adequate information about the temperature of the room
- Deterministic? ... **No**
  - A door / window may be open
- Episodic? ... **No**
  - AC is continuously working for it's lifetime
- Static or dynamic? ... **Dynamic**
  - The environment can change even without any action of the AC
- Discrete or Continuous? ... **Discrete**
  - Though temperature of the room can vary continuously, we model it as a finite set $\{OK, HOT\}$

Purely reactive agent

Agent without memory



- Agent has no memory
  - Does not remember history / experience
- Action is determined by the current environment state alone
- $action : S \to A$

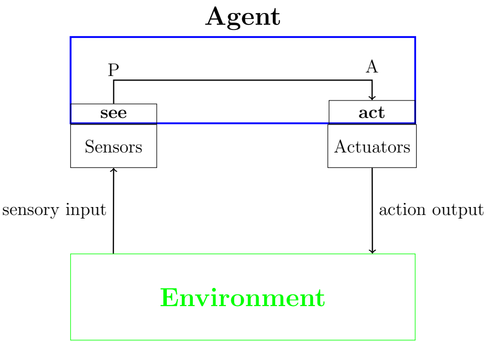
Purely reactive agent

Example: A Room AC

- Environment states: $S = \{OK, HOT\}$
- Actions: $A = \{START, STOP\}$
- Agent behavior: $action : S \to A$
  - $action : OK \to STOP$
  - $action : HOT \to START$
- Environment model: $env : S \times A \to \rho(S)$
  - ... Independent of agent behavior
  - $env : OK, START \to \{OK, HOT\}$
  - $env : OK, STOP \to \{OK, HOT\}$
  - $env : HOT, START \to \{OK, HOT\}$
  - $env : HOT, STOP \to \{HOT\}$
    - We assume that it is not snowing outside! ☺

Modeling Perception

Purely reactive agent



- "see" maps an environment state to a percept: $see(s) = p$
  - $see : S \rightarrow P$
- If $s_1 \neq s_2$, but $see(s_1) = see(s_2)$, i.e. $p_1 = p_2$
  - Environment states s_1 and s_2 are indistinguishable for the agent
  - ... e.g., there can be a person in the room that the AC cannot see
- We revise our stand
  - Agent behavior is determined by percept
  - $action : P \rightarrow A \quad$ (NOT $S \rightarrow A$)

Purely reactive agent with perception

Example: A Room AC

- Person in room: Y
- Environment states: $S = \{(OK, Y), (OK, !Y), (HOT, Y), (HOT, !Y)\}$
- AC has a thermostat only – cannot see if somebody is in the room
- Percepts: $S \to P$
  - $see(OK, Y) \to ok$
  - $see(OK, !Y) \to ok$
  - $see(HOT, Y) \to hot$
  - $see(HOT, !Y) \to hot$
- The states (OK, Y) and $(OK, !Y)$ are indistinguishable, both lead to percept ok
  - Similarly, for (HOT, Y) and $(HOT, !Y)$

Perception for room AC

Continued

- Agent behavior: $action : P \rightarrow A$ ** not $S \rightarrow A$
  - $action : ok \rightarrow STOP$
  - $action : hot \rightarrow START$
- Environment model: $env : S \times A \rightarrow \rho(S)$
  - $env : (OK, Y), START \rightarrow \{(OK, Y), (HOT, Y), (OK, !Y), (HOT, !Y)\}$
  - $env : (OK, !Y), START \rightarrow \{(OK, Y), (HOT, Y), (OK, !Y), (HOT, !Y)\}$
  - ...

An AC with a camera

… In addition to a thermometer

- Environment states: ** Same as earlier
  - $S = \{(OK, Y), (OK, !Y), (HOT, Y), (HOT, !Y)\}$
- Percepts: $S \rightarrow P$
  - $see(OK, Y) \rightarrow ok, y$
  - $see(OK, !Y) \rightarrow ok, !y$
  - $see(HOT, Y) \rightarrow hot, y$
  - $see(HOT, !Y) \rightarrow hot, !y$

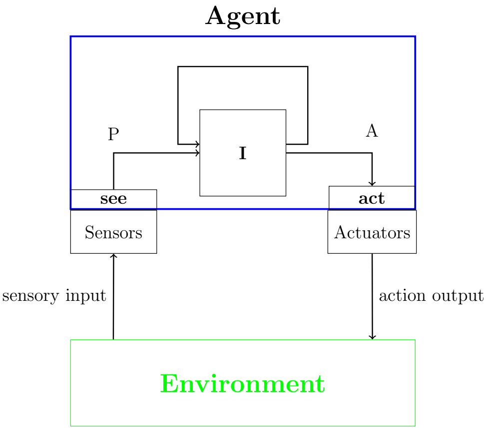
A room AC with a camera

Continued

- The AC should switch on only if the AC <u>sees</u> a person in the room and <u>perceives</u> the temperature to be hot
- Agent behavior: $action : P \rightarrow A$
  - $action : ok, y \rightarrow STOP$
  - $action : ok, !y \rightarrow STOP$
  - $action : hot, y \rightarrow START$
  - $action : hot, !y \rightarrow STOP$

- Environment model: $env : S \times A \rightarrow \rho(S)$ ** Same as earlier
  - $env : (OK, Y), START \rightarrow \{(OK, Y), (HOT, Y), (OK, !Y), (HOT, !Y)\}$
  - $env : (OK, !Y), START \rightarrow \{(OK, Y), (HOT, Y), (OK, !Y), (HOT, !Y)\}$
  - ...

Reactive agent with memory

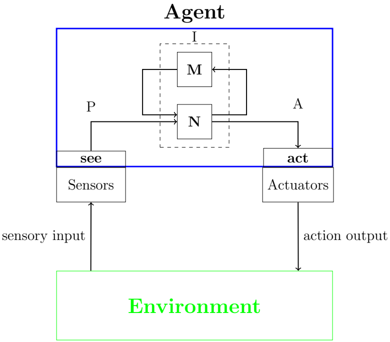


- $see : S \to P$
- $next : P \times I \to I$
- $action : I \to A$
- I stores the history
  - Memory is finite
  - How many steps?
  - Abstracted form?

Reactive agent with memory

Example: A Room AC (no camera) with a constraint

- If it is already ON, it cannot START, and vice-versa

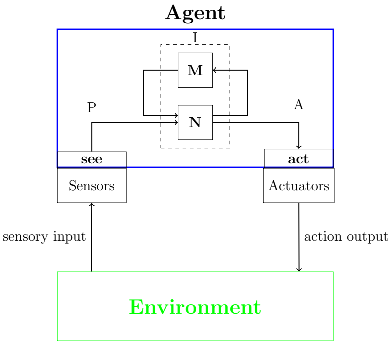


- $next : P \times M \to N$
- $mem : N \to M$, $action : N \to A$

- Percepts: $P = \{ok, hot\}$
- Actions:
  $A = \{START, STOP, NOP\}$
- Memory: $M = \{ON, OFF\}$
- Agent behavior $next : P \times M \to N$
  - $next(ok, ON) = (STOP, OFF)$
  - $next(ok, OFF) = (NOP, OFF)$
  - $next(hot, ON) = (NOP, ON)$
  - $next(hot, OFF) = (START, ON)$

Example: A Room AC (no camera) with a constraint

... Continued



- $next : P \times M \to N$
- $mem : N \to M$, $action : N \to A$

- Memory: $N \to M$
  - $mem(STOP, OFF) \to OFF$
  - $mem(NOP, OFF) \to OFF$
  - $mem(NOP, ON) \to ON$
  - $mem(START, ON) \to ON$
- Action: $N \to A$
  - $action(STOP, OFF) \to OFF$
  - $action(NOP, OFF) \to NOP$
  - $action(NOP, ON) \to NOP$
  - $action(START, ON) \to ON$

Can we model a Reactive Agent as a table look up?

Example: A Room AC

- Agent behavior: $N = (A, M)$
  - $next(ok, ON) = (STOP, OFF)$
  - $next(ok, OFF) = (NOP, OFF)$
  - $next(hot, ON) = (NOP, ON)$
  - $next(hot, OFF) = (START, ON)$
- Table lookup for *next*

	ok	hot
ON	$STOP, OFF$	NOP, ON
OFF	NOP, OFF	$START, ON$

- Theoretically possible, but may need extremely large tables

Quiz

Quiz 06-03

End of Module 06-03

Biological Vision and Applications
Module 06-04: Deliberative Agents

Hiranmay Ghosh

Reactive vs. goal-directed behavior

- The AC reacts to the perceived environment states
  - Does not have an explicit representation of goal
  - It's goal (to maintain the room temperature) is implicit in the agent behavior
  - Reactive behavior
- Contrast it with an autonomous car
  - It has an explicit goal: to reach a destination
    - In shortest time, via a scenic route, etc.
  - Proactive / goal-directed behavior needs deliberation
- A practical agent should show both reactive and proactive behavior
  - It should take you to destination ... safely
  - Apply brakes on seeing a pedestrian at front

Reactive vs. Deliberative Agents

What is deliberation

Human mental process: a model for goal-directed behavior

- Deliberation is the act of thinking about or discussing something and deciding carefully (Merriam-Webster dictionary)
  - “I take an umbrella **because** I see that it is raining” (perception)
  - “I give you money **because** I believe that you deserve it” (belief)
  - “I take the school bus **because** I want to go to school” (desire)
- Reasoning happens with percepts, beliefs, desires
- A more complex deliberation process
  - I take taxi **because**
    - I want to go to cinema, and
    - I believe that taxi will take me to cinema, and
    - I see that it is raining
- Deliberation (reasoning) is a property of human mind

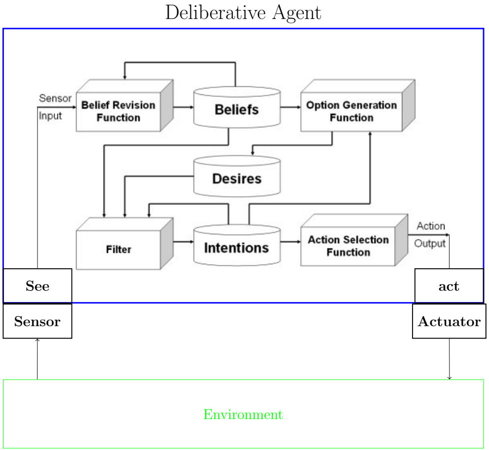
Modeling Deliberative Agents: BDI Architecture

The internal state of the agent is modeled

- **Belief:** An agent's model of the world
  - There is a difference between belief and knowledge
  - Updated with perceptions, communication (additional information)
- **Desire:** The ideal state that an agent wants to accomplish
  - The world is not in that state at the current moment
  - A desire might be realistic or not
  - Relatively static, may change in long-term
- **Intention:** A subgoal required to fulfil the desires
  - Determines it's current activity

Belief vs. knowledge

BDI Architecture



Belief Revision Function:
$brf : B \times P \to B$

Option generation function:
$options : B \times I \to D$

A filter function [Deliberation]:
$filter : B \times D \times I \to I$

Action selection function:
$action : I \to A$

Example of deliberation in an autonomous agent

- Autonomous car: goal is to reach a destination by 3pm
- Desire: Want to reach a destination by 3pm
- Belief: A map (road network)
- Percept: Current position and orientation
- Intention: Find a feasible path
- Action: Do planning (graph search), with temporal constraints

Example of deliberation in an autonomous agent

Contd.

- Desire: Want to reach a destination by 3pm
- Belief: A map, best path to destination $(X \rightarrow Y \rightarrow Z \ldots)$
- Percept: Current position and orientation
- Intention: take road X
- Action: navigate the car

Example of deliberation in an autonomous agent

Contd.

- After sometime, we are on road Y
- Desire: Want to reach a destination by a certain time
- Belief: A map, best path to destination $(X \rightarrow Y \rightarrow Z \ldots)$
- Percept: Current position and orientation, road-block ahead
- Intention: Current commitment (road Y), Recompute feasible path
- Action: Replan
  1. Alternative found: accept alternate path
  2. Alternative not found: change desire

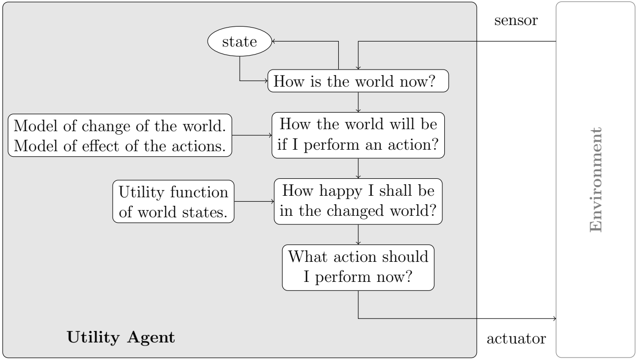
Preferential order of states

- The desire of an agent is to achieve an "ideal" state for itself
  - ... an agent prefers certain states to others
- Let $W = \{w^0, w^1, \ldots w^n\}$ represents the set of the states of the world
- Preferential ordering:
  - $w_i \succsim_A w_j$ means that the agent A either prefers state w_i to state w_j or is indifferent to them
  - $w_i \succ_A w_j$ means that A strictly prefers w_i over w_j by agent A
  - $w_i \approx_A w_j$ means that A is indifferent to w_i and w_j
- Important: Preferential order of states are specific to agents
- Can the preferential states be partially ordered?

Utility of states

- Utility is a mapping from the world states to a numeric value by an agent A
  - $u_A : W \to \mathbb{R}$
- Consistent with the preferential order of the states for the agent
  - $u_A(w_i) \geq u_A(w_j) \iff w_i \succsim_A w_j$
  - $u_A(w_i) > u_A(w_j) \iff w_i \succ_A w_j$
  - $u_A(w_i) = u_A(w_j) \iff w_i \approx_A w_j$
- Utility values of states are specific to agents
- Rational behavior implies that the agent tries to maximize it's utility

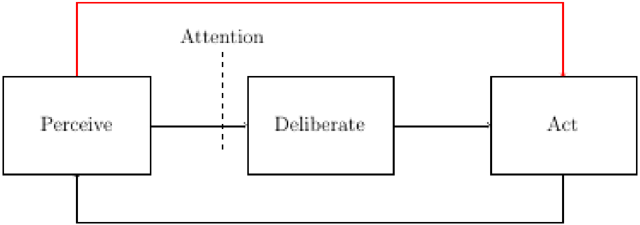
Architecture of a utility based agent



Agent Model

Key Takeaway

- A perpetual loop for it's lifetime



- Reactive behavior is a behavior without deliberation
- Deliberation is the key to goal-directed behavior
- "Mind" of the agent is responsible for the deliberation
  - Mind is modeled with memory, and associated processing

Quiz

Quiz 06-04

End of Module 06-04

Biological Vision and Applications
Module 06-05: Symbolic & Emergent Systems

Hiranmay Ghosh

Computational modeling of cognition

Two competing paradigms

- Symbolic approach
  - Motivated by AI research (symbolic logic)
    - Examples: Logic systems
  - Knowledge-driven (top-down) approach
- Emergent system approach
  - Motivated by neurosciences (neural networks)
    - Examples: Classifiers, Neural networks
  - Data-driven (bottom-up) approach

Symbolic approach is traditionally known as the “cognitive approach”

Comparison of symbolic and emergent system approaches

Knowledge representation & sharing

- Symbolic approach:
  - Explicit and formal representation of knowledge
    - Constructed from atomic symbols and their relations (model)
  - Humans can understand / contribute
    - Symbolic systems generally start with human implanted knowledge
    - Can work from day 1 (without training)
  - Knowledge can be easily shared across systems
    - Multiple agents can work with a common shared knowledge-base

- Emergent system approach:
  - Knowledge implicit in the states of the system (no explicit model)
  - Cannot be understood / contributed by humans
    - Knowledge is acquired through training (from data)
  - Knowledge is strictly private to the system
    - Can be shared only through explicit communication

Comparison of symbolic and emergent system approaches

Knowledge creation, maintenance & update

- Symbolic approach:
  - can be created / maintained by humans
    - Can be engineered in a modular way (knowledge partitions)
  - Representation is static
    - Update requires external intervention
    - Periodic knowledge update

- Emergent system approach:
  - Knowledge is implicit in the states of the system
  - Monolithic: Cannot be partitioned
    - Cannot be engineered
  - Knowledge gets updated in real-time with interaction with the environment
    - Continuous knowledge update

Comparison of symbolic and emergent system approaches

Generality and flexibility

- Symbolic approach:
  - Constructed from atomic symbols and their relations
  - Can combinatorially generalize (inductive generalization)
    - Experience "red apple" and "yellow banana"
    - Conceive the concepts of "yellow apple" and "red banana"
  - Lacks flexibility – cannot discover new concepts
- Emergent system approach:
  - Monolithic: cannot generalize
  - Flexible
    - Can find new / shifting patterns in data

Comparison of symbolic and emergent system approaches

Mode of reasoning

- Symbolic approach:
  - Deliberative
    - Follows a formal process
  - More reliable: Accuracy can be estimated
  - Less tolerance to noisy data: brittle
  - Explainable
  - Slow
- Emergent system approach:
  - Intuitive
    - No formal process
  - Less reliable: Accuracy cannot be estimated
  - More tolerance to noisy data: robust
  - Not explainable
  - Fast

See table 6.1 in book

How does the human mind work?

Dual process theory

- Symbolic approach: formal, knowledge driven, explainable, accurate, slow
  - More suitable for deliberations
- Emergent system approach: informal, data-driven, not explainable, inaccurate, fast
  - More suitable for reactive behavior

- Dual process theory: Human mind uses both the approaches
- How do they interact ?
  - Parallel model: fast and slow thinking occur simultaneously
    - ... and may conflict
  - Default-Interventionist model: fast thinking generates intuitive responses
    - Subsequent slow thinking processing may or may not deliberate on them

Dual process theory ...

How does the cognitive systems work?

- Most of cognitive systems till date are based on symbolic approach
  - ... Emergent approach is catching up fast
- Each approach has it's own advantages and disadvantages
  - Symbolic approach is more suitable for cognitive tasks
  - Emergent approach is more suitable for perceptual data processing
- A fusion between the two can be the key to build large practical cognitive systems
  - Bayesian reasoning is an amalgamation of knowledge-driven and data-driven reasoning
  - "Graph Networks" is yet another approach where neural networks use prior models

Quiz

Quiz 06-05

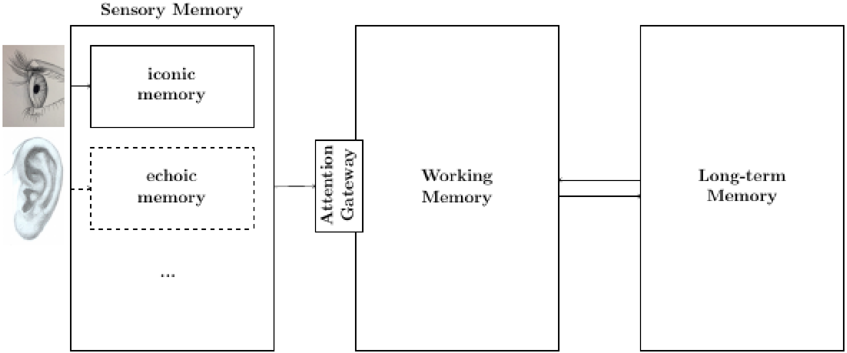
End of Module 06-05

IIT Jodhpur

Biological Vision and Applications
Module 06-06: Short Term Memory

Hiranmay Ghosh

Memory pipeline for perception & cognition



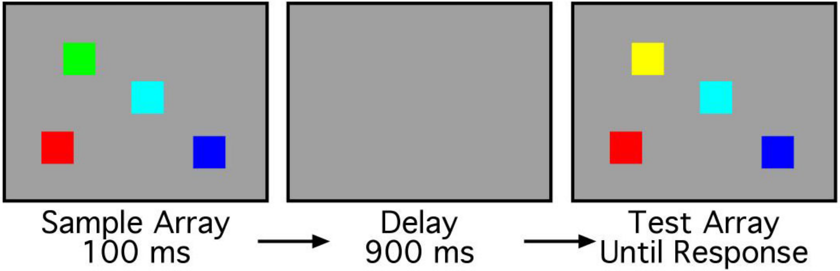
- Memory is the key enabler for deliberation and learning in cognitive systems

Iconic Memory

- Holds visual information as icons
- Pre-attentive stage of vision
- Extremely short retention period
  - ≈ 100 ms
- Coherent representation of
  - Entire visual perception
  - For a very short time
- Leads to persistence of vision / motion blur
- The duration of visible persistence is inversely related to
  - Stimulus duration
  - Stimulus luminance

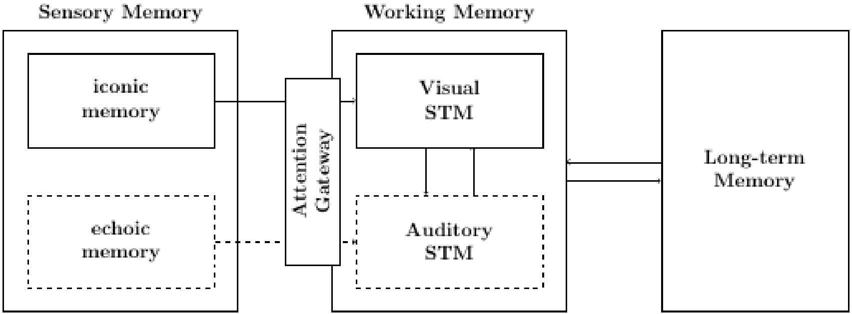

;

Iconic Memory & Change blindness



- The delay erases the memory

Working Memory / Short Term Memory



- Independence of visual and verbal tasks
- Further subdivisions in ASTM / VSTM ?

Working Memory vs Short Term Memory

- Short term Memory
  - Remembering something for a short while
    - ... e.g. a telephone number
- Working memory
  - Manipulating the contents of the memory
    - ... e.g. mental maths (adding two numbers)
- Interaction between STM and WM
  - Recall the numbers to add
  - Remember longer with mental rehearsal
- What is the representation of information in STM / WM ?

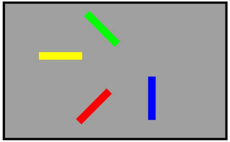
Visual Short Term Memory (VSTM)

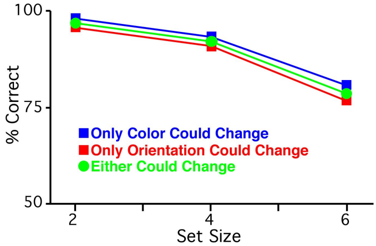
- Retention period longer than iconic memory
  - Few seconds
  - Stores both “what” and “where”
  - Bridges sensory gap over saccades
- Rehearsal increases retention period
  - Repetition without interpretation (understanding)
  - In symbolic form ?
  - ... sends information to LTM ?
  - (rote learning)

Scholarpedia article

Capacity of VSTM

- Small capacity: 3 – 4 "items"
  - Varied over individuals / experiments
  - Limit on volume of information, or number of objects ?
- How do we experience the plethora of objects around us ?
  - Just-in-time acquisition of needed information
- Multi-feature objects
  - Integrated object representation is stored
  - Justifies object-based attention

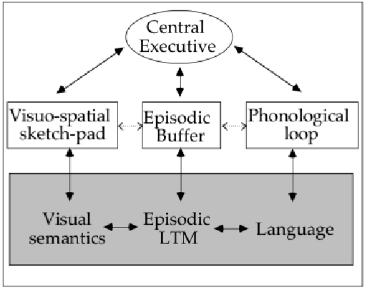




Vogel. Storage of features, conjunctions, ... in visual working memory (2001)

Working memory

- Accessible for deliberative (goal-directed) information processing
- Central executive
  - Exercises attentional control over other components
- Visuo-spatial sketchpad
  - Create and maintain task-specific visual images
- Episodic buffer
  - Holds multi-modal information
  - Combines task-specific information from VSTM, ASTM and LTM



Scholarpedia article

Quiz

Quiz 06-06

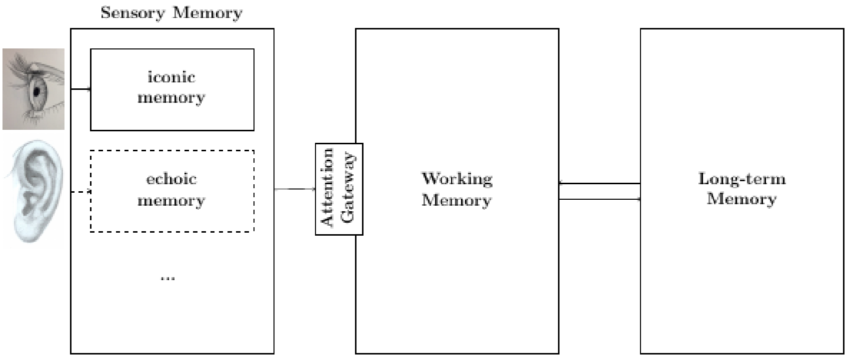
End of Module 06-06

IIT Jodhpur

Biological Vision and Applications
Module 06-07: Long Term Memory

Hiranmay Ghosh

Memory pipeline



Long-term and short-term memory

- Short-term memory (STM)
  - Stores sensory information, current context (internal)
  - Can recall data from Long term memory
  - Used for deliberation
- Long-term memory (LTM)
  - Stores information for a longer term
    - ... indefinitely ?
  - Stores knowledge, experience
  - Unlimited capacity
    - Accessibility (recall) can be a constraint

Long-term memory (LTM)

- **Declarative Memory (Explicit)**: *What*
  - Stores declarative (explicit) form of knowledge
  - Recalled consciously and reasoned with
    - A banana is yellow
    - An apple is round
    - Sampled marbles in bag no. 1 are all blue
- **Procedural Memory (Implicit)**: *How*
  - Stores the (implicit) knowledge about how to solve a problem
    - A machine learned image classifier
  - Knowledge is not recalled consciuosly – cannot be explicitly described

Declarative Memory

- Semantic:
  - General facts that universally hold good
    - A banana is yellow
  - Independent of personal experience
- Episodic:
  - Memory of previous experiences, including
    - context (time, place, associated events, emotions)
  - Can be recalled after it has happened
    - Sampled marbles in bag no. 1 are all blue
- Episodic knowledge may be abstracted to Semantic knowledge
  - All marbles in bag no. 1 are blue
  - All bags contain marbles of same color

Retention in Long-term Memory

Is LTM "permanent"?

- Sometimes we fail to recall experience / knowledge
  - Decay in episodic memory (permanent)
  - Retrieval failure (temporary – some time, can recover with cues)
- Decay
  - We generally tend to remember experiences that evoke extreme emotions longer
    - ... Not always true
  - LSTM is an attempt to implement selective retention of episodic memory
  - Modeling decay in semantic memory is even more complex
- Retrieval failure – difficult to model

Associative Memory

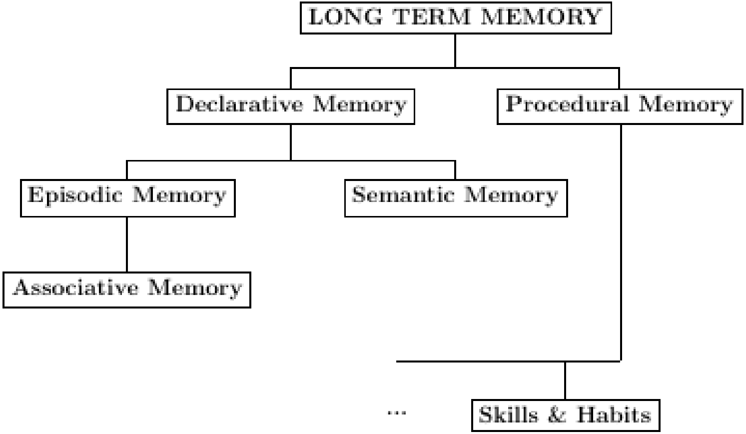
A class of Episodic Memory

- Particularly useful for sensory perceptions
- Associates entities (e.g. objects) with sensory properties (e.g. shape, color, ...)
- The object can be quickly recalled from the perception
- Also known as Content Addressable Memory (CAM)
- Why is associative memory episodic (not semantic)?

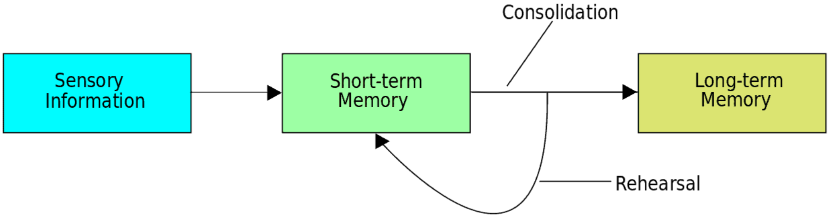
Procedural Memory

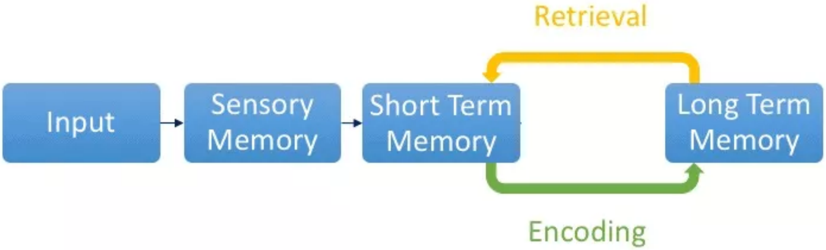
- Can be hold various types of procedural information
  - Skills and habits (how-to's)
    - Example: How to recognize an object
    - Implemented as a classifier (black-box)
    - Done this way by an agent ... could be done in a different way too
  - ...

Long-tem Memory



STM and LTM


Consolidation
Sensory Information
Short-term Memory
Long-term Memory
Rehearsal


Retrieval
Input
Sensory Memory
Short Term Memory
Long Term Memory
Encoding

Quiz

Quiz 06-07

End of Module 06-07

Biological Vision and Applications
Module 06-08: Global Workspace Theory

Hiranmay Ghosh

What is consciousness ?

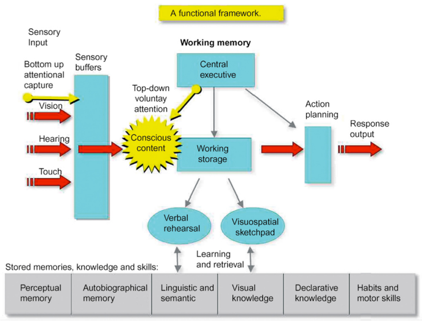
- Consciousness is everything that you “experience”
  - We do not experience all that we perceive
  - There are many facors that results in “attention” to specific stimulii
    - bottom-up and top-down
    - environment, context, intention, ...

Dual Process Theory

Conscious and Subconscious Processing

- You are conscious about the squirrel
- In subconsciousness, you are also looking at the tree, ...
- Consciousness is dynamic
  - Suddenly, a bird flies out of the tree
  - ... your consciousness will shift there

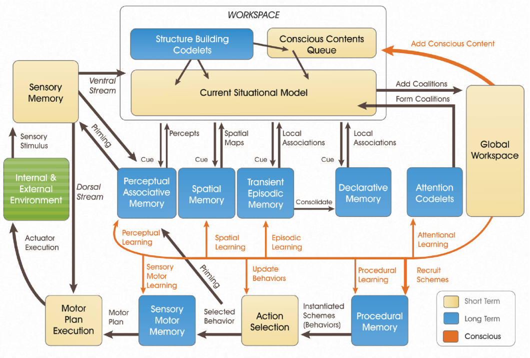
Global Workspace Theory



- "Theatre" metaphore
- Many independent processes work simultaneously on the memory
- Sensory stimulii and actions are the actors on the stage
- Context puts the spotlight on a part of the memory
- Conscious events enable learning
  - explicit & implicit
- Unconscious resources controls the conscious activities

Global Workspace Theory ...

LIDA Architecture



- LIDA Cognitive Architecture (Video)
- LIDA Tutorial

Quiz

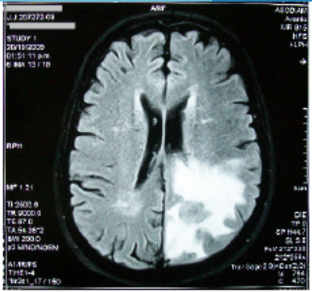
No quiz for module 06-08

End of Module 06-08

Biological Vision and Applications
Module 07-01: Knowledge Representation

Hiranmay Ghosh

Knowledge required for visual interpretation



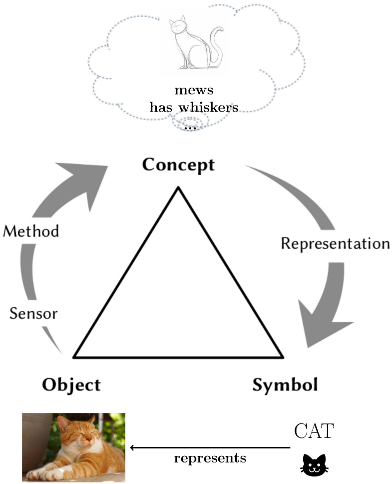
- Types of knowledge required
  - Domain Knowledge (about anatomy / astronomy)
    - also called ontology
  - Knowledge about image formation / processing
    - Mapping of real objects to images ... and vice-versa
    - How to interpret an image
  - The relation between the two

Characterizing knowledge

- Domain knowledge:
  - Declarative: explicit and symbolic representation
  - exists independent of processing structure
  - can be shared
- Knowledge about image formation / processing:
  - Procedural: implicit
  - Encoded as algorithms, neural networks or classifiers
  - Strictly private to the processing scheme
- We shall focus on declarative knowledge in this module

Symbolic representation

The semiotic triangle



- Objects (things): That exist
- Concepts: Mental representations (models)
- Representation: Symbol to represent a concept (text, icon, audio)

Representational Theory of Mind (RTM)

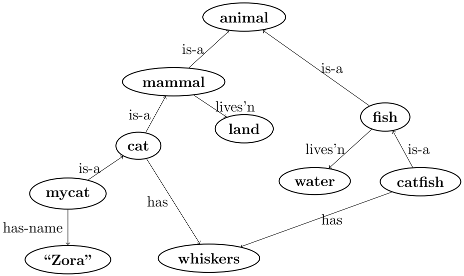
- A concept is a mental model of “something” that exists (with attributes)
- Something can be
  - A real-world thing
  - An internal mental state of the agent
- A name is associated with a concept
  - For reference during manipulation (reasoning)
- Knowledge is
  - A collection of named concepts
  - A set of sentences (propositions) that relate the concepts
    - Named concepts: cat, tail, has
    - Proposition: A cat has a tail

Language of Thought Hypothesis (LoTH)

- Thoughts are mental processes
  - Leads to mental models
  - Result of manipulation of the knowledge
    - That brown cat has a tail
    - If I had wings!
- Represented in a language that is akin to symbolic logic
- Inferences can be drawn from knowledge through the process of thought (reasoning)

Semantic Network

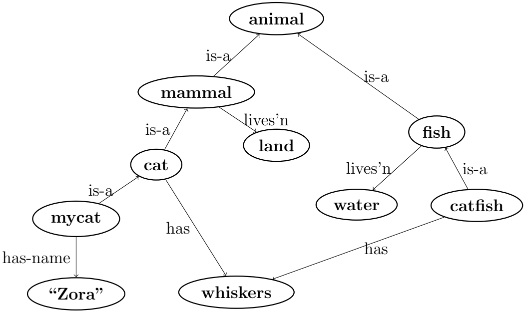
- Knowledge is a set of statements
  - A mammal is an animal
  - A cat is a mammal
  - A cat has whiskers
  - A mammal lives on land
  - A fish is an animal
  - A catfish is a fish
  - A catfish has whiskers
  - A fish lives in water
  - Mycat is a cat
  - Mycat has a name "Zora"
- Equivalently, knowledge is a graph (semantic network)



Semantics of “Semantic Network”

- Each edge of a semantic network represents a proposition (statement)
- Each proposition describes a property of a concept
- For example: cat has whiskers
  - Subject (Concept being described): Cat
  - Predicate (Property): Has
  - Object (Value): Whiskers
- A concept can be a class, or an instance
- A value can be a concept, or a literal
- The network of concepts represent knowledge about a domain

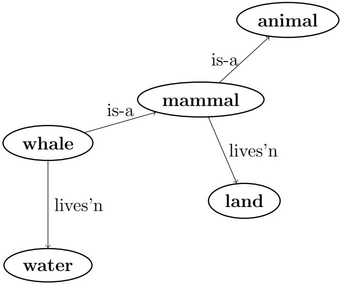
Reasoning with Semantic Network



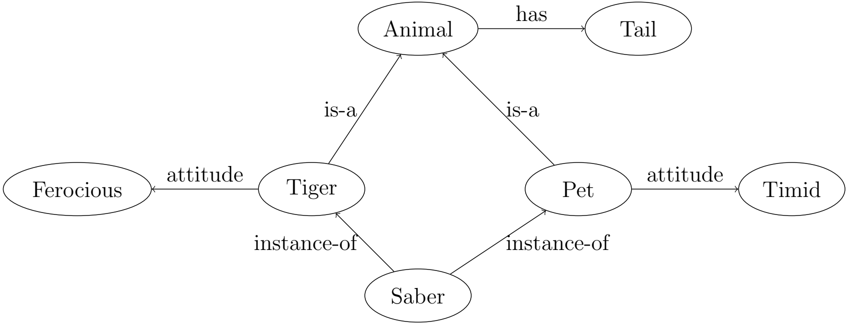
- Requires underlying axioms, e.g.
  - Properety inheritance
    - If *a* "is-a" *b*, then *a* inherits properties of *b*
  - "is-a" is transitive
    - If *a* is-a *b*, and *b* is-a *c*, then *a* is a *c* too
  - These axioms make a semantic network efficient (compact)

Flexibility with "Semantic Network"

- No restrictions on properties / values to be associated to a concept
- There can be exceptions. e.g.
  - Whale is a mammal, but lives in water
- Axioms need to be redefined
  - If *a* "is-a" *b*, then
    - *a* inherits properties of *b*
    - ... unless overruled



Multiple inheritance



Saber has a tail

What's about Saber's attitude ?

Properties of Semantic Network

- A semantic network is extremely flexible
  - At the cost of formalism
  - An informal description of a domain (in it's basic form)
- Semantics is imposed with axioms / constraints
- Many variants have been proposed
  - Definitional network
    - Expresses class-subclass relations
    - ... and properties that distinguish sibling subclasses
    - Cat *is-a* mammal; cat *has* whiskers
  - Implication Network
    - Expresses causal relations
    - Banana *causes* yellow color
  - Hybrid networks combine more than one of paradigms

Sowa. Semantic Networks

Resource Description Framework (RDF)

- A knowledge representation framework based on semantic network
- All entities are treated as "resources"
  - Each resource is identified with a IRI
  - Enables distributed knowledge description
- An RDF sentence is a triplet $\langle \mathit{subject}, \mathit{predicate}, \mathit{object} \rangle$
- A predicate in one sentence can be a subject or an object in another
  - ⟨ hasWeightInKg, is-a, healthParameter ⟩. ⟨ Ramu, hasWeightInKg, 80 ⟩
- Reification: A statement is also a resource (and identified by an IRI)
  - I said that cat is an animal
  - S1: ⟨ cat, is-a, animal ⟩. S2: ⟨ I, said, S1 ⟩
- Constraints and semantics defined with RDF and RDF Schema
- Notations: XML, N3, Turtle

SPARQL Query Language

- To make query on RDF Graphs
- Syntactically similar to SQL
- Implemented with "triple-store" databases
  - Apache Jena / TDB
  - Optimized for storing triplets
- Query on distributed knowledge
  - Distributed knowledge centrally indexed
  - Distributed query processing (distributed index)

- Resources:
  - W3School tutorials
  - W3C Documents

Quiz

No quiz for module 07-01

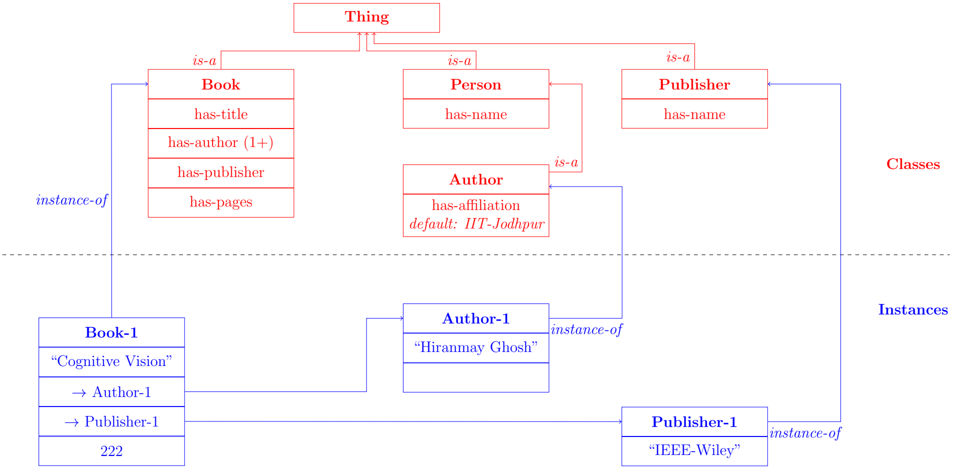
End of Module 07-01

IIT Jodhpur

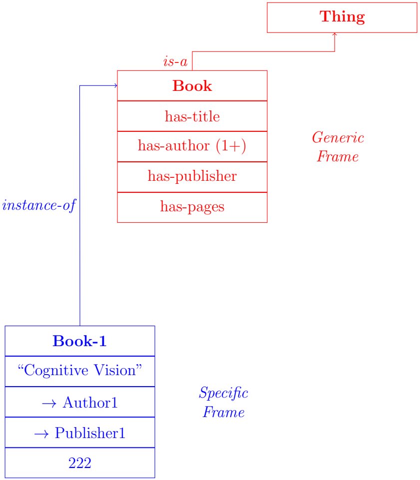
Biological Vision and Applications
Module 07-02: Frame-based representation

Hiranmay Ghosh

Frame-based representation



Frames, slots and fillers

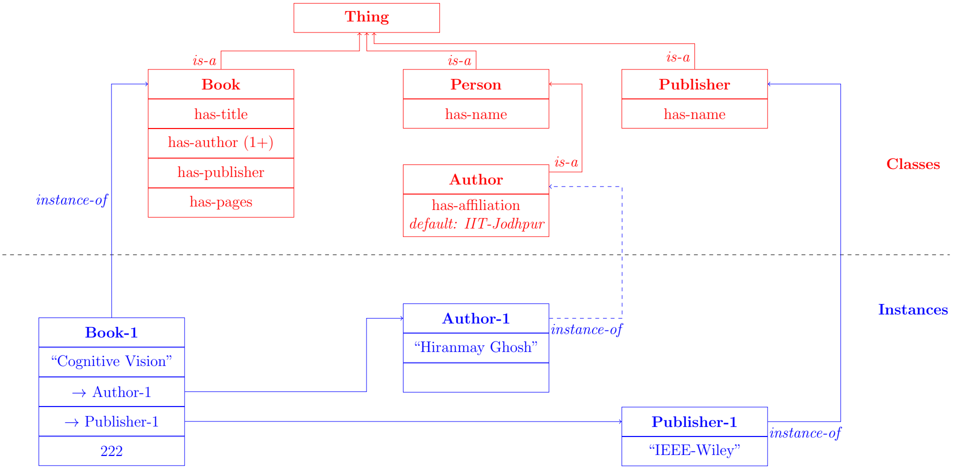


- A frame has a label
- A frame consists of one or more slots (attributes)
- A slot contains a filler (value)
  - Reference to another frame
  - Literal
  - not specified
- A frame inherits attributes and default values of it's parent
- Value restrictions
  - Data types / range
  - Cardinality

Ontology and data

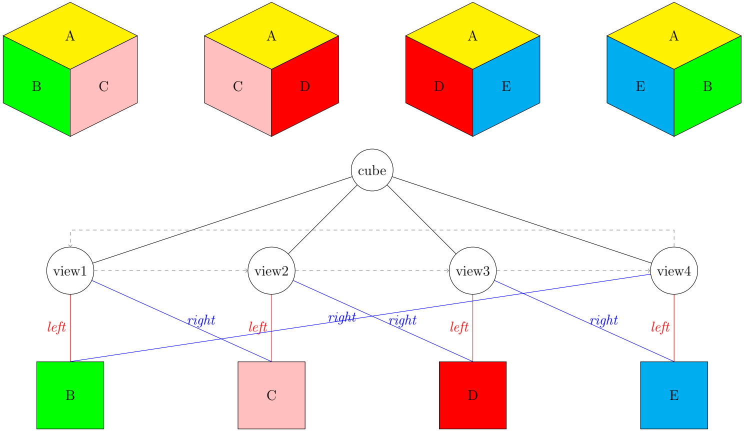
- The generic frames and their interconnections define a model (schema) for a domain
  - A domain is a bounded part of the world
  - The model is also known as an ontology
  - An ontology imposes constraints on data and their organization
- The specific frames represent instances of the classes (data)
  - They are defined and organized following the constraints of the ontology
- Web Ontology Language (OWL)
  - W3C recommended standard for web-based knowledge representation
  - Is defined as a schema over RDF/RDFS

Inferencing with frame-based representation



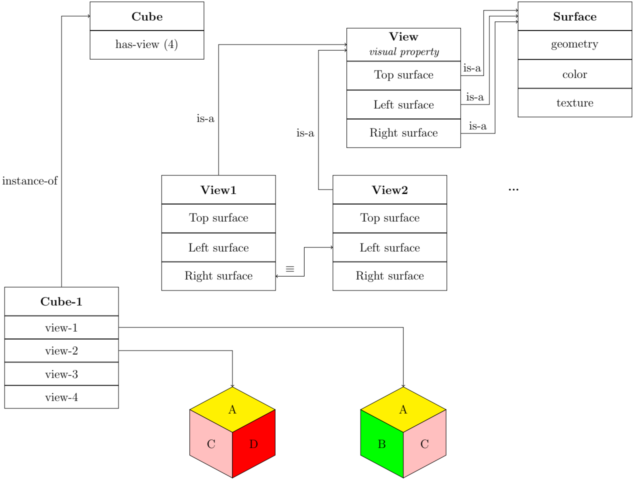
Frame based representation & Visual cognition

Visual events and viewpoints

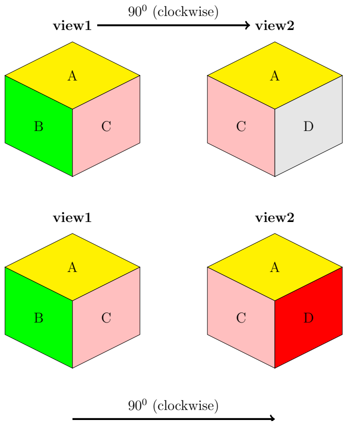


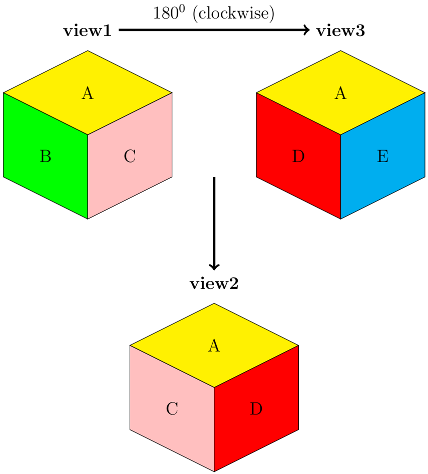
Frame based representation & Visual cognition

Viewpoints: frame-based representation

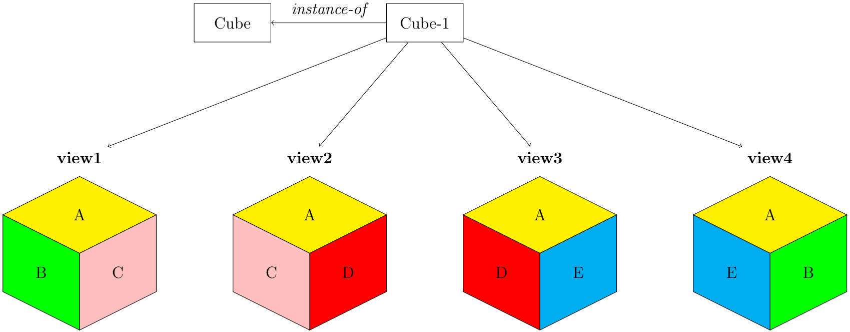


Inferencing with visual frames

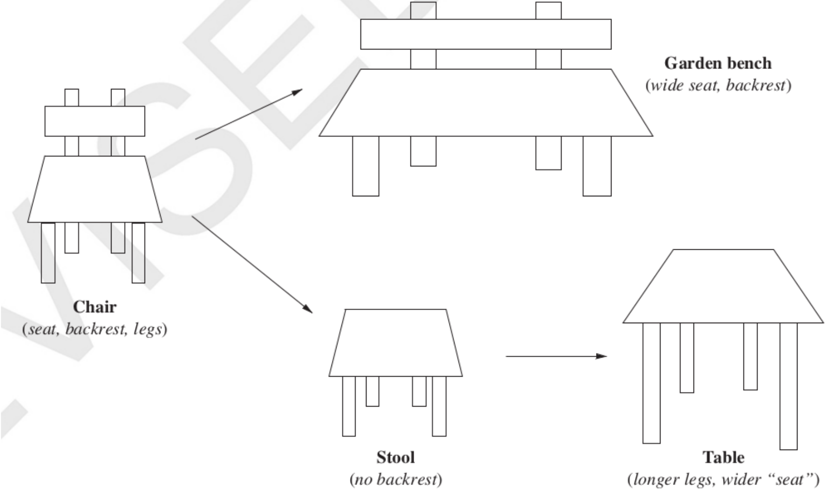




Inferencing with visual frames



Specialization of concepts



Frame-based representation and Visual cognition

Summary

- A compact and efficient representation of visual world
- A frame represents a specific view of a system (object / scene)
  - Remembered in declarative memory
- A frame-system is a collection of frames representing different views of a system
  - Different frames of a system describe the system from different viewpoints
  - Change of viewpoint (movement) results in transformations across the frames
- When one receives a new percept, one recalls the nearest matching frame from memory
  - Leads to object recognition
- If no available frame sufficiently match the current situation, the closest frame is extended to define a new system

Minsky's paper (1974)

Quiz

Quiz 07-02

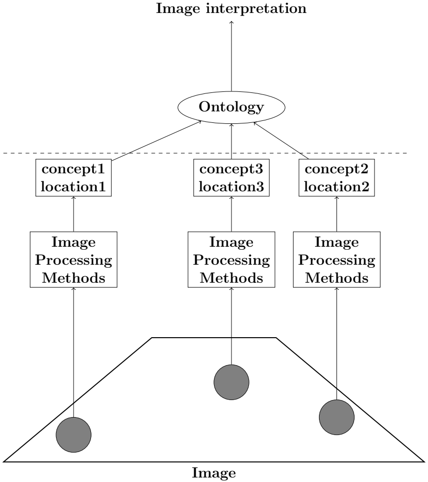
End of Module 07-02

IIT Jodhpur

Biological Vision and Applications
Module 07-03: Knowledge representation for visual cognition

Hiranmay Ghosh

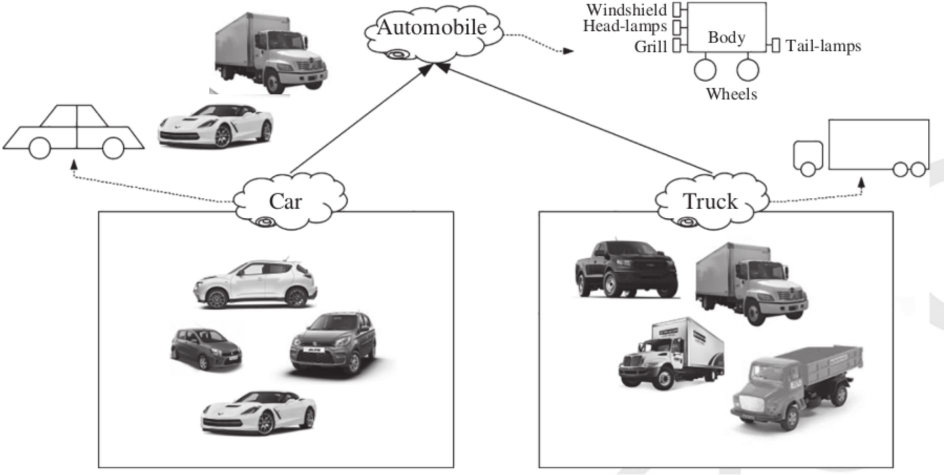
Traditional practice in ontology-based computer vision



- Media data and concepts are processed separately and differently
- Context is missing for image processing Semantic Gap
-
  - Many-to-many mapping between objects and their visual features
  - Often resolved with context

How concepts are formed

… Taxonomy learning



How to build perceptual ontology

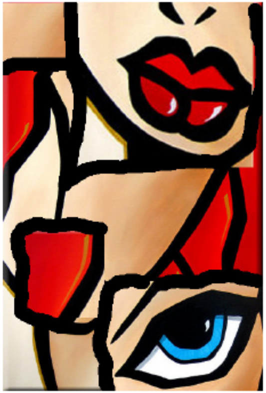
- Include perceptual description (declarative) of an object in the ontology
  - ball – color – red (symbolic)
  - Limited capability for feature representation

- Matching problem starts when we put numeric feature descriptors
  - e.g. color value expressed as r-g-b
  - how do we compare the colors in the logical framework?
  - crisp yes/no answer does not work

- Need probabilistic matching
  - Merger of Bayesian Network and Knowledge graph
    - Banana has yellow color with probability 0.7
    - $\langle\langle \mathit{Banana}, \mathit{color}, \mathit{yellow}\rangle, \mathit{prob}, 0.7\rangle$

Baye's OWL

Spatial / Temporal Structure

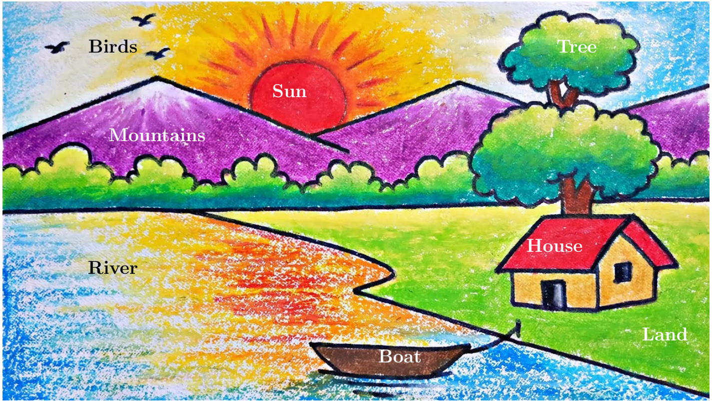


Structural composition of a scene

Example

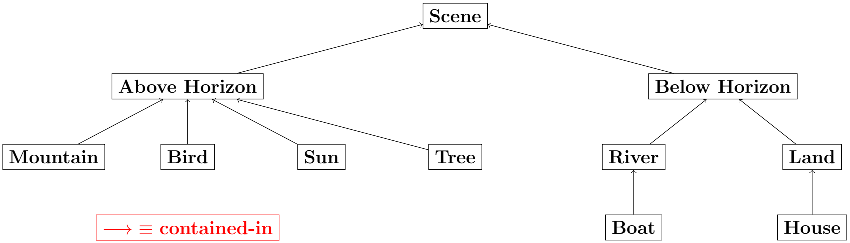
Above
Horizon

Below
Horizon



Scene hierarchy

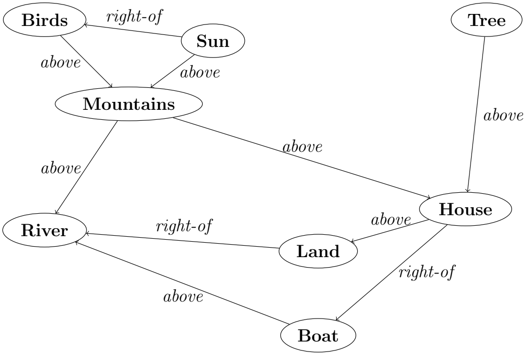
Describes hierarchical composition of a scene



- Reasoning: A contained-in B, B contained-in C $\rightarrow$ A contained-in C

Scene graph

Describes network structure of a scene



- Reasoning:
  - A above B, B above C $\rightarrow$ A above C
  - Inverse relations: inverse(right-of) = left-of, etc.

Encoding knowledge about a scene

Coping up with intrinsic variations

- Avoid unnecessary details
  - Do not overspecify
- Express as qualitative relations
  - Approximate relations
- Express the relations as probably distribution

Graph matching (short paper)

Quiz

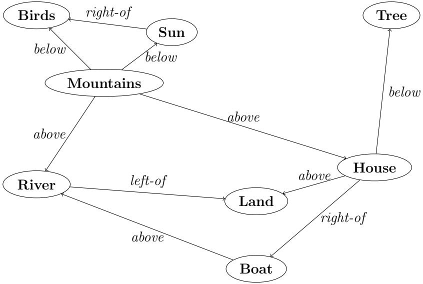
No quiz for module 07-03

End of Module 07-03

Biological Vision and Applications
Module 07-04: Spatio-temporal relations

Hiranmay Ghosh

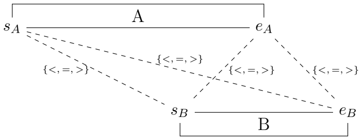
Informal (normative) relations



- The relations "left-of", "above", etc. are informal
  - ... lacks semantics
- How do we formally specify them ?

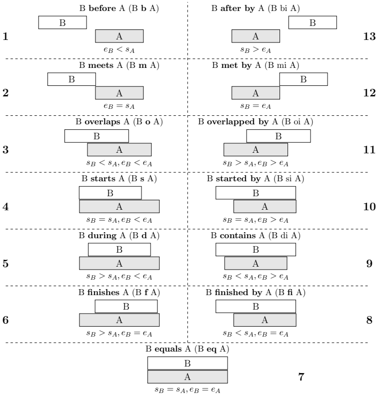
Allen's temporal relations

- An event A spans a finite interval of time
  - Start and end points: s_A, e_A
  - Finite and positive duration: $s_A < e_A$
- Two point events x and y can be has three possible unambiguous relations
  - $x < y$, $x = y$ and $x > y$
- Temporal relation between two interval events A and B can be represented as
  - Comparison 4-tuple of $(s_A, e_A) \times (s_B, e_B)$
  - Are there 3^4 possible values ?



Allen's temporal relations

13 feasible distinct relations

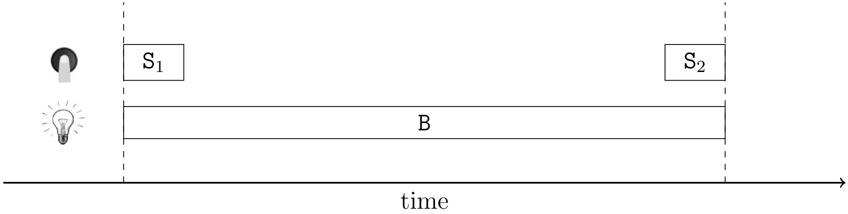


1. $e_B < s_A$: `B` *before* `A`
2. $e_B = s_A$: `B` *meets* `A`
3. $s_B < s_A, e_B < e_A$: `B` *overlaps* `A`
4. $s_B = s_A, e_B > e_A$: `B` *starts* `A`
5. $s_B > s_A, e_B < e_A$: `B` *during* `A`
6. $s_B > s_A, e_B = e_A$: `B` *finishes* `A`
7. $s_B = s_A, e_B = e_A$: `B` *equals* `A`

Allen's temporal relations

Example

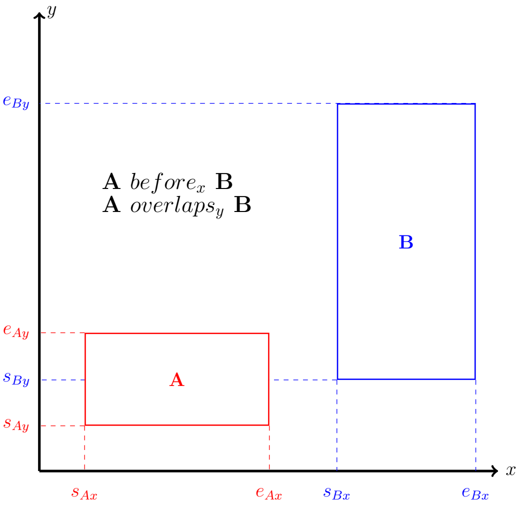
(a) $\mathtt{S}_1$ *before* $\mathtt{S}_2$
(b) $\mathtt{S}_1$ *starts* $\mathtt{B}$
(c) $\mathtt{S}_2$ *finishes* $\mathtt{B}$



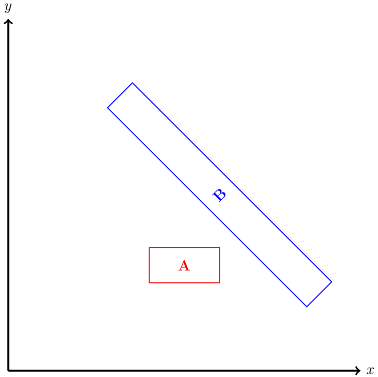
Allen's relations

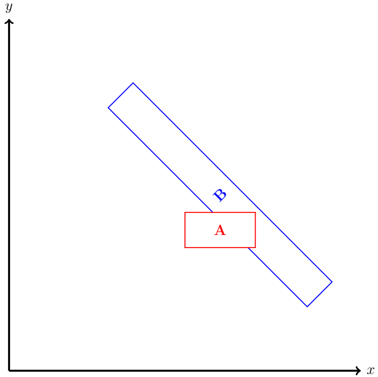
Extension to spatial dimensions

- Can be applied to spatial dimensions as well
  - "before" $\rightarrow$ "left-of" / "below"
- Express spatio-temporal relations as a tuple of allen relations
  - (A b_x B, A o_y B)



Ambiguity: Allen's relations extended to multi-dimensional space

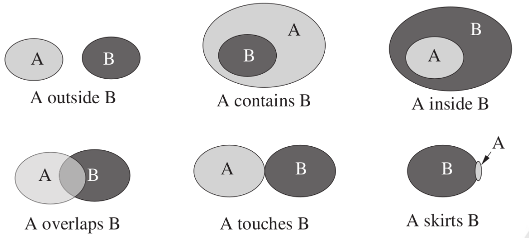




- In both the cases, (`A` d_x `B`, `A` d_y `B`)
  - Left: A does not intersect B
  - Right: A intersects B

Containment relations (multi-dimensional)

To resolve ambiguity



- In multi-dimensional space
  - Spatio-temporal relations unambiguously defines with
    1. The Allen's relations on projections on each axis
    2. The containment relations (in multiple dimension)

Quiz

No quiz for module 07-04

End of Module 07-04

IIT Jodhpur

Biological Vision and Applications
Module 07-05: Qualitative spatial and temporal relations

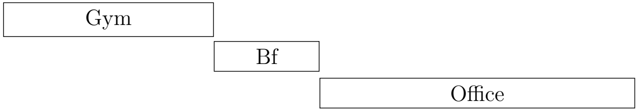
Hiranmay Ghosh

Allen's Interval Algebra

- Consider three 1D interval events: A, B, C
  1. A r_{AB} B
  2. B r_{BC} C

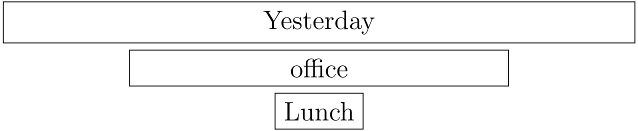
- Can we infer the relation between A and C ?
  - A r_{AC} C
  - Given r_{AB}, r_{BC}, can we find r_{AC} ?

An Intuitive Introduction



- Consider the statements:
  1. I went to gym just before having my breakfast: Gym m Bf
  2. I went to office immediately after the breakfast: Bf m Office
- We can conclude
  - Temporal relation between Gym and Office: Gym b Office
- Transition rule: $(m, m) \rightarrow b$
- **An example of temporal sequencing**

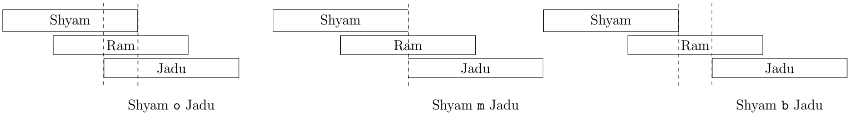
An Intuitive Introduction (contd)



- Consider the statements:
  1. I attended office some time during yesterday: Office d Yday
  2. I ate my lunch while at office: Lunch d Office
- We can conclude
  - Temporal relation between Lunch and Yesterday: Lunch d Yday
- Transition rule: $(d, d) \rightarrow d$
- **An example of hierarchical decomposition**

An Intuitive Introduction (contd)

1. Ram came in the room while Shyam was there and continued after Shyam left:
   - Shyam o Ram
2. Jadu came into the room when Ram was there and continued after Ram left:
   - Ram o Jadu



- The temporal relation between Shyam and Jadu cannot be uniquely resolved
  - Shyam b, m, o Jadu
- Transition rule: $(o, o) \rightarrow \{b, m, o\}$

Allen's temporal algebra

- Given that `A` r_{AB} `B` and `B` r_{BC} `C`
    1. where r_{XY} is one of the Allen's relation
- Temporal constraint between `A` and `C`: `A` R_{AC} `C`
    - where R_{AC} is a subset of Allen's relation
- Mapping $r \times r \xrightarrow{T} R$ is defined over a transitivity table
    - $R \leftarrow T(r_1, r_2)$
    - 13×13 entries in transitivity table

Allen's transitivity table

Allen's Interval Algebra

Generalizing ...

- R_{ij}, R_{jk}, R_{ik}: Temporal constraints between event-pairs
- $(\mathtt{E}_i, \mathtt{E}_j)$, $(\mathtt{E}_j, \mathtt{E}_k)$ and $(\mathtt{E}_i, \mathtt{E}_k)$
  - In general, each is a subset of Allen's relations
- The algorithm for computing $\mathit{Constraint}(R_{ij}, R_{jk}) \neq R_{ik}$:

Algorithm 18: Computing relational constraint

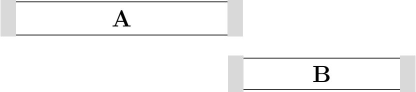
```
procedure Constraint(R_ij, R_jk)
    C = ∅;
    for each p ∈ R_ij do
        for each q ∈ R_jk do
            C ← C ∪ T(p,q);
        end
    end
    return C;
end
```

Uncertainty with the endpoints

- Where do the shore ends and the sea starts

- A man is walking – He falls
- When does he start falling and when does he end falling ?

- Measurement error

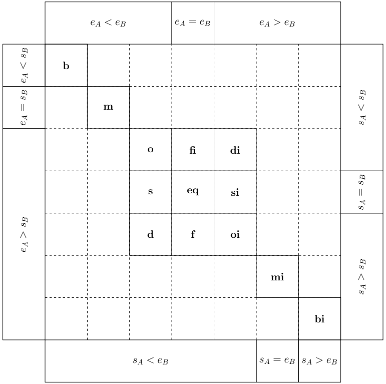


A *before* **B** ?
A *meets* **B** ?
A *overlaps* **B** ?

Approximate qualitative relations

Conceptual neighborhood

- Relations organized in 2D defines conceptual neighbors
- Ambiguity in boundary / Measurement error may lead to a relation to be misclassified in it's conceptual neighborhood
- A set of relations in a conceptual neighborhood defines an "approximate relation"
- For fuzzy representation, see book



Semantics of Allen's relations

Quiz

No quiz for module 07-05

End of Module 07-05

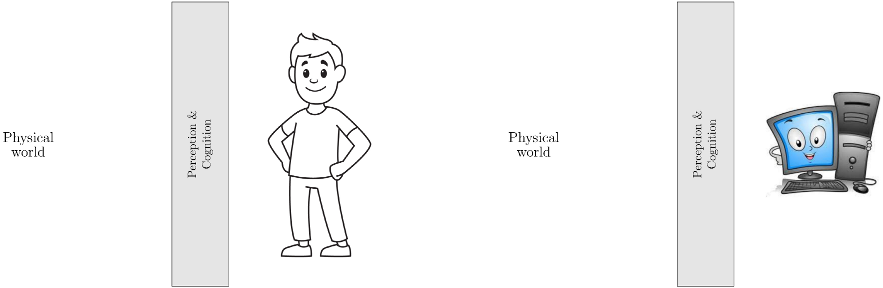
Biological Vision and Applications
Module 08-01: Applications

Hiranmay Ghosh

Application areas

- We have reviewed 13 papers (2005 – 2019)
  - Visual Query Answering
  - Semantic Labeling
  - Content Adaptation
  - Content Recommendation
  - Surveillance
  - Photo Enhancement
  - Image Restoration
  - Scene Reconstruction
  - Human Robot Interaction

Common thread



- Transform visual signals as humans would perceive them
- Decide what is important – what should be processed and what should be ignored
- Context and human intention
- Fast and intuitive (hard real-time for some applications)

Guiding Principle 1

Recognize "Semantic Gap"

- Visual signals (features) and "semantics" do not correspond to each other
- Need for in-context abstraction of media features
  - Video summarization
    - Foreground-background separation
    - "Semantic" features (face, music, ...)
  - Painting Restoration
    - Cracks vs. lines
- The most difficult challenge in computer vision

Guiding principle 2

Contrast conveys information

- Principle of early vision
  - Contrasts: Color (R-G, B-Y) & illumination
  - Edge detection & perceptual grouping
  - Natural Scene Statistics
- Used in initial feature extraction / preprocessing
  - Crack detection (Painting, pipeline, railway track, ...)
  - Model of quality & aesthetics
    - Computational photography
- Convolution is a universal tool

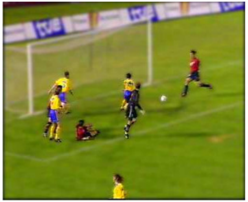
Guiding principle 3

Contextual Semantics is conveyed through a very small fraction of the scene

- Decide what is important in a given context
  - Identification of important concepts in visual contents (robotics)
  - Fast & real-time processing (surveillance)
  - Semantic labeling and (visual) query answering
  - Video compression (storage & transmission)
    - "Signal-level fidelity" vs. "Semantic fidelity"
- Drastic reduction in information precessing
- Principle of attention is crucial for cognitive vision

‘

Glas, et al. 2012



Cavallaro, et al. 2005

Guiding principle 4

Use contextual information

- Context disambiguates
- Context can be found elsewhere – not in the image alone
  - Caption (image / video)
  - Metadata (date/time, camera parameters, ...)
  - Markers in the environment
- Applications:
  - Semantic labelling & VQA
  - Robotics

He & Hu, 2019

Guiding principle 5

Principle of Inductive Generalization

- Apply knowledge from one task to another
  - Transfer learning
  - Few-shot or one-shot learning
  - Zero-shot learning
  - Multi-task learning
- Methods
  - Use of structured knowledge (machine learned)
  - Hierarchical Bayesian Model
- Applications:
  - Dealing with rare concepts / new queries
  - Cross-recommendation
  - Face region detection
  - Surveillance (railway track monitoring)

More examples

Other principles

- 6. Use of emergent knowledge
  - Where reliable models do not exist / difficult to codify
  - Examples: aesthetics, cracks in paintings
  - ML techniques (Clustering, Neural networks, ...)
- 7. Global workspace
  - Multiple processes working in parallel
  - Cognition to action (robotics)
- 8. Social networking
  - Learn from each other
  - Collaborative learning (robotics)
    - Imitation learning

Summary

Key takeaways from this course

- "Cognitive vision" encompassed all computer vision tasks
  - There are no specific applications of cognitive vision
  - Application of principles of biological vision in computer vision tasks
- There is no unified theory / framework for Cognitive Vision yet
  - Each topic covered in the course is an isolated dot
  - Cognitive Vision is like a vast ocean – We have explored some islands in the ocean

Your presentations, participations & reviews

Some common improvement suggestions

- Presentation
  - The key points (with respect to our class) need to be brought out clearly
  - Need to go beyond the paper (important methods used / critical thoughts)
  - Verbatim reading from the slides/notes makes a presentation drab
  - Need to generate enough interest in the audience – discussions
  - Time management – neither too short, not too long
- Participation & Review
  - Interactions required during/after presentation
  - Summary should be of optimal lenght – need to bring out key points crisply
  - Need professional assessment – dispersion of awarded marks
  - Specific comments (strengths/weaknesses)
  - Ratings should be adequately justified – especially very high and low ones
    - "Could not follow the presentation", but awarded high scores

Quiz

No quiz for module 08-01

End of Module 08-01

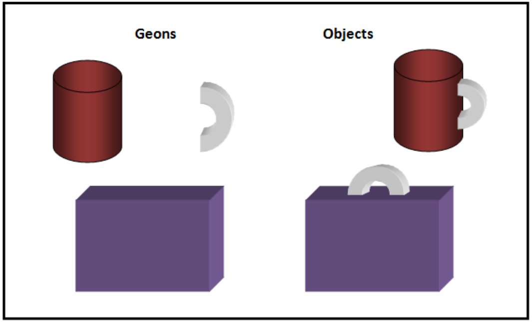
IIT Jodhpur

Biological Vision and Applications
Module 09-01: Graph Neural Networks

Hiranmay Ghosh

Structured representation

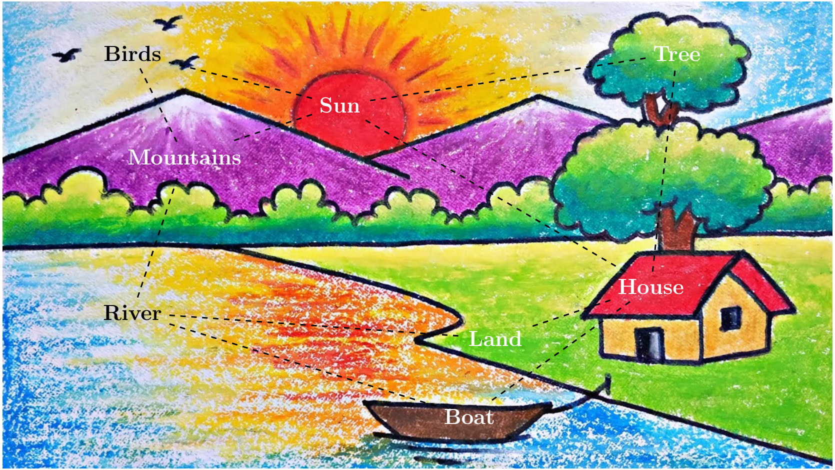
Explicit knowledge



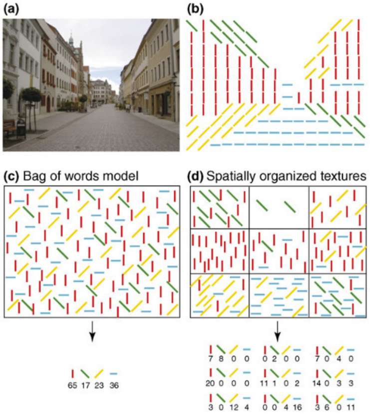
- Inductive Generalization
- Context

Structured representation

Spatial (and temporal) Organization



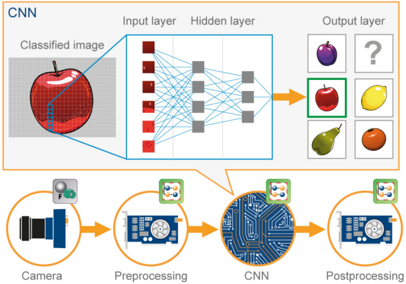
Holistic representation



- **No Inductive Generalization**

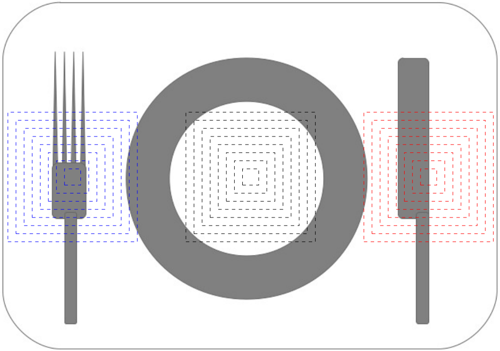
Machine learning

Holistic representation



- Does not use the explicit structured representation
- Flexibility: Knowledge is emergent
  - No dependence on hand-coded knowledge
  - Features and feature weights are machine learned

Does a CNN "see" the structure ?



Can we combine the benefits of the two approaches ?

Position paper from DeepMind, Google Brain, MIT, University of Edinburgh (2018)

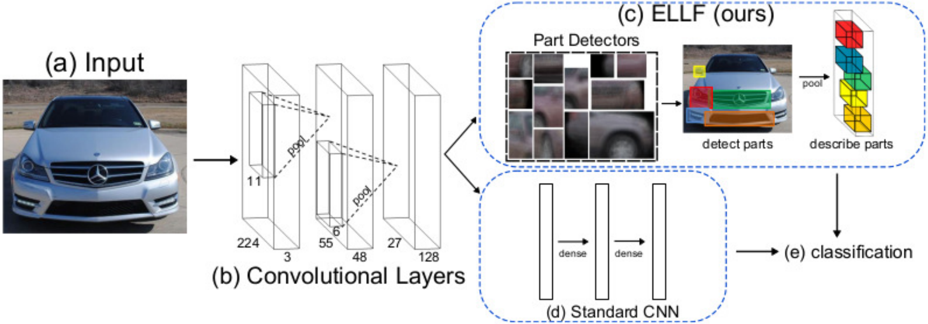
- Key to advancement of AI (cognition): Inductive generalization
- Structure with flexibility (emergent knowledge)
- "Intersection of deep learning and structured approaches"
  - Reason (following DL approach) on structured data (expressed as graph)

- Indeed, this realization is not new!

Battaglia, et al. Relational inductive biases, deep learning, and graph networks (2018)

Fine-grained object recognition

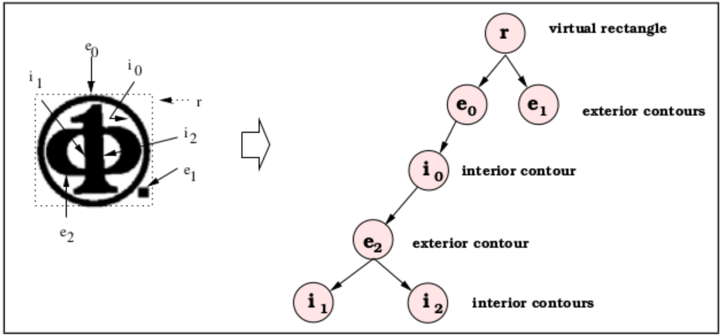
Car model (ICPR'14), Bird species (CVPR'15), ...



- Attention for discriminative parts

Recursive Neural Network (1998)

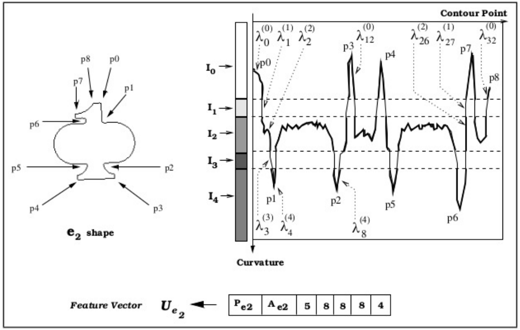
Example: Logo recognition



- Identify the external and internal contours by image processing techniques
  - Edge detection, perceptual grouping
- Create a tree structure

Recursive Neural Network

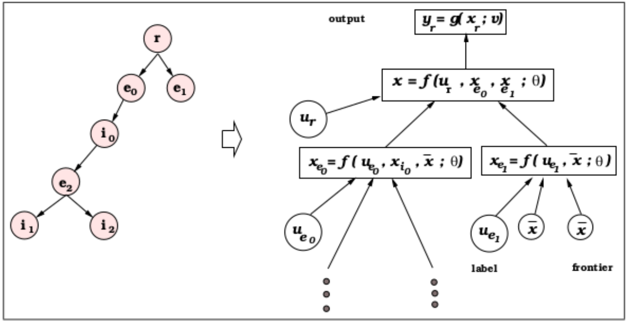
Feature representation



- Shape descriptor for each contour:
  - Perimeter, area
  - Histogram of curvatures

Recursive Neural Network

Processing model



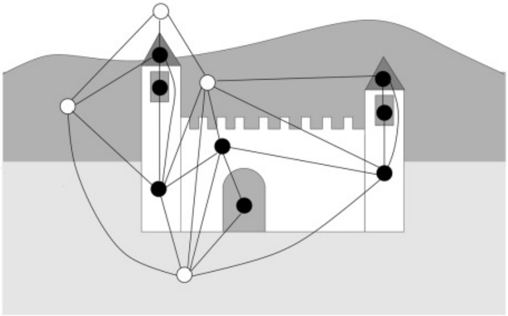
- Functions f() and g() can be realized as deep neural networks
- Identical property descriptions u_x

Recursive Neural Network

Training

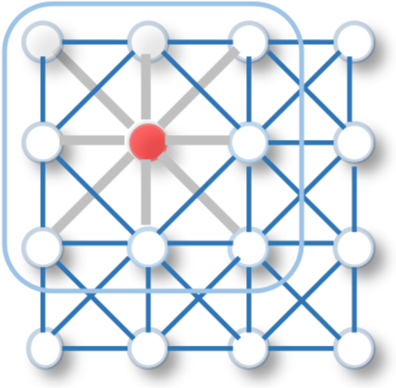
- Trained over a large logo database
  - Each logo generates a different tree
- Goal is to learn parameters for f() and g()
  - Same f() and g() for every node / tree
- Better accuracy than MLP based approach
  - Exploits structure information
  - Parameters (features / weights) are machine learned
- Computations at "Lower" nodes affect that at "higher" nodes, not vice-versa
- Only graph-level (global) inference is drawn
- Frasconi, et al. A General Framework for Adaptive Processing of Data Structures
- Frasesconi, et al. Logo Recognition by Recursive Neural Networks

Graph Neural Network

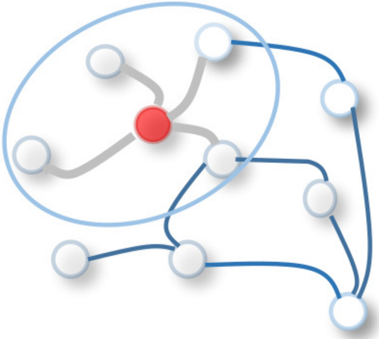


- Find super-pixels in the image
- Create a graph
  - Nodes: Super-pixels
  - Edges: Adjacent nodes
- Inferencing
  - Graph focussed: Castle
  - Node focussed: Tower, Door, Window, ..., Background

2D Convolution vs. Graph Convolution



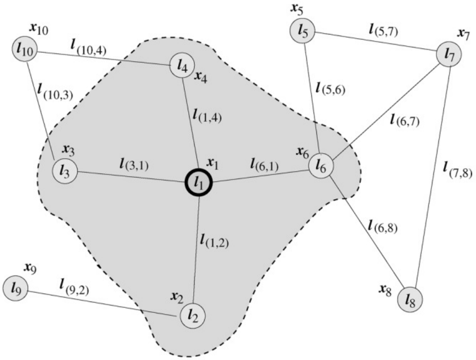
2D Convolution



Graph Convolution

Convolutional Graph Neural Network

Processing model

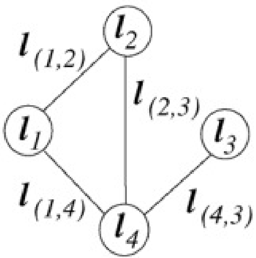


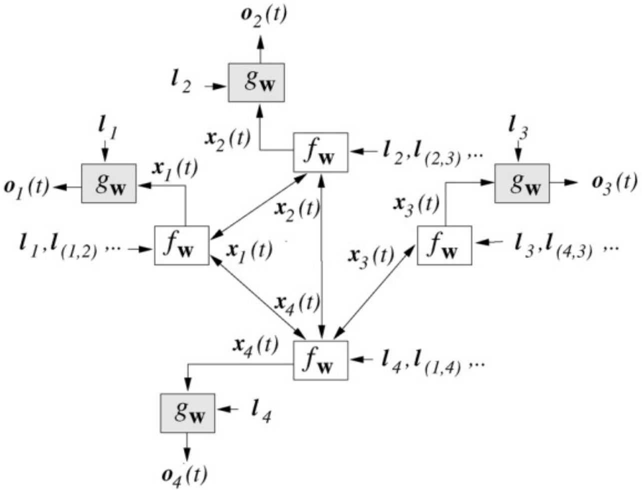
$$x_1 = f_w(l_1, \underbrace{l_{(1,2)}, l_{(3,1)}, l_{(1,4)}, l_{(6,1)}}_{l_{co[1]}}, \underbrace{x_2, x_3, x_4, x_6}_{x_{ne[1]}}, \underbrace{l_2, l_3, l_4, l_6}_{l_{ne[n]}})$$

- Nodes have identical property (feature) descriptors
  - e.g. color, texture, shape
- Edges have identical property descriptors
  - e.g. distance between the center of gravities of the nodes

Graph Neural Network

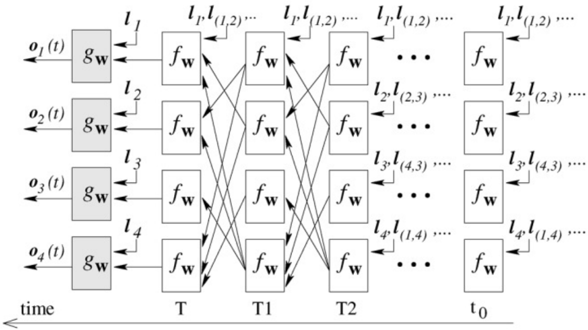
Processing model (contd.)





Graph Neural Network

Recurrent Processing

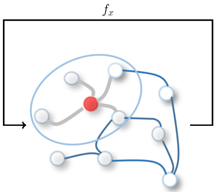


- Take the output after several recursions
  - Is the system guaranteed to go into a steady state after a finite number of iterations?

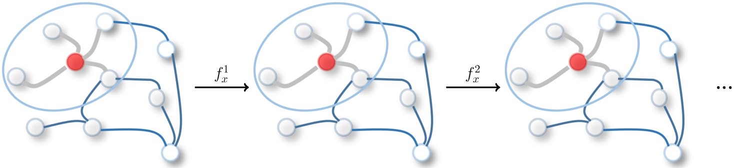
Scarselli, et al. The graph neural network model (2009)

Recursive & Convolutional Graph Neural Network

Rec-GNN & Conv-GNN



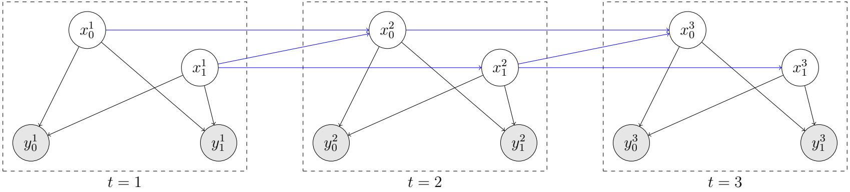
Recursive GNN



Convolutional GNN

Temporal Dependencies

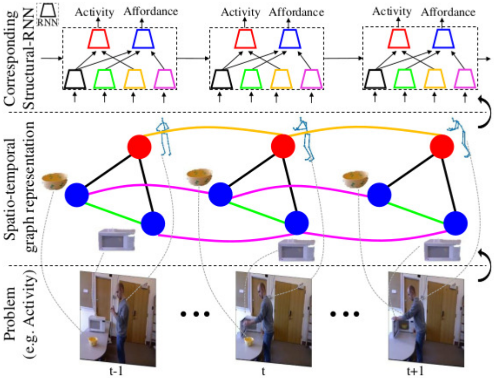
Dynamic Bayesian Network



- Network parameters: $P(X^0), P(Y^t \mid X^t), P(X^t \mid X^{t-1})$
- Prob distribution at $t = T$:
  - $P(X^T) = P(X^0).\prod_{t=1:T} P(Y^t \mid X^t).\prod_{t=1:T} P(X^t \mid X^{t-1})$

Spatio-Temporal Graph Neural Network

Structured RNN (S-RNN)



- CVPR-16 Paper, Presentation Video

Applications

- Rec-GNN / Conv-GNN (still images)
  - Fine-grained classification (e.g. bird species)
  - 3D point cloud processing (LiDAR)
- Spatio-Temporal GNN (video / motion picture)
  - Human action recognition
  - Human/Robot - Object Interaction
  - Human Motion Modeling

Limitations and Future Research

- Model depth
  - Performance (accuracy) drops with depth
- Scalability
  - Graph size needs to be limited
  - Number of nodes / number of edges
- Heterogeneity of graphs
  - Presently graphs are assumed to be homogeneous
- Dynamicity
  - Graph structure changing over time

Quiz

Your feedback for the course please

End of Module 09-01

IIT Jodhpur

Biological Vision and Applications
Module 03-01: Reasoning Paradigms

Hiranmay Ghosh

What is "reasoning"

- We "know" some facts
  - Supplied by others
  - Sensed by some sensors (percept)
- We infer unknown facts from the known facts

A simple example:

Premises

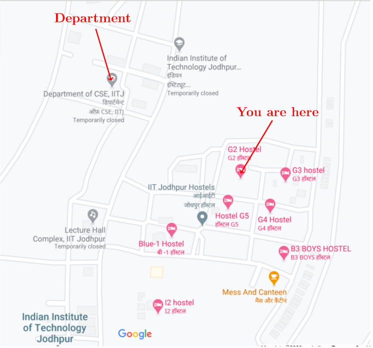
1. I am at Bangalore
2. My college is in Jodhpur
3. Train no 16508 goes to Jodhpur

$\Rightarrow$

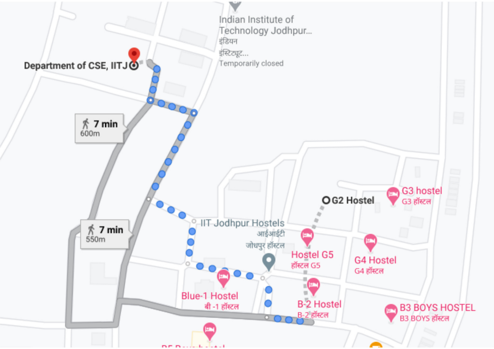
Inference

I can take train no 16508 to go to my college

One more example



Premises

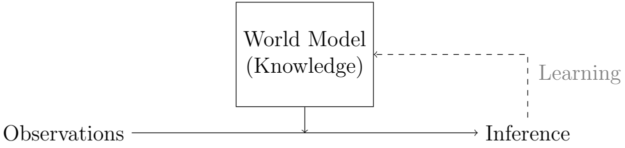


Inference

Reasoning paradigms

- In human mind, reasoning is intuitive
- For application to machines, we need to formalize the algorithms
- Reasoning paradigms
  - Knowledge driven (top-down)
    - Model based
    - Case based
  - Data driven (bottom-up)

Model-based reasoning



- Create a model of the world of discourse (knowledge)
- Interpret observations with that model leading to inference
- Learning: Inference may lead to change in the model

> Model-based reasoning is the formal way of interpreting observations with the model

Rule-based reasoning

Also called deductive reasoning

Kowledge / Model:
$\forall x, parrot(x) \rightarrow fly(x)$

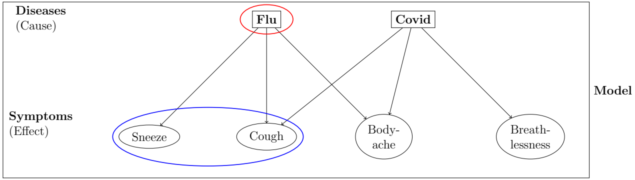
Observation:
$parrot(mybird)$

Inference:
$fly(mybird)$

- Formalized as logic
  - Identical representation for knowledge, observation and inference (statements)
  - Define some formal rules
  - Apply an appropriate rules on knowledge + observations to “deduce” new statements
- Many flavors
  - Propositional calculus, predicate calculus
  - First order logic, second order logic, ...
  - Descriptions logic

See Norvig & Russell

Abductive reasoning



What is the best expnation for the observations ?

- Inexact match – robustness
  - Model may not be accurate – lack of knowledge
  - Inherent system uncertainty
  - Observations may be missing / inaccurate

Comparing deductive and abductive reasoning

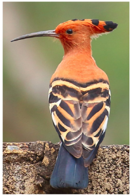
- Reasoning is valid in deductive reasoning
  - If the premises are true, the consequence must be true – can be proved.
  - Inference may not always be correct for abductive reasoning
- Deductive reasoning can discover facts implied by known facts only
  - Abductive reasoning can discover new facts
  - ... e.g., detecting a new human face
- Deductive reasoning needs accurate information on premises
  - If premises are not accurately known, the reasoning breaks down
  - Abductive reasoning is robust

Induction

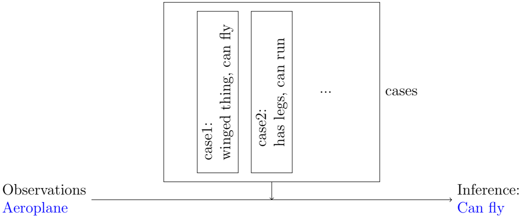
Generalization from observations

- Example: Suppose you observe
  - Parrot is a bird; parrot can fly
  - Crow is a bird; crow can fly
  - Mynah is a bird; mynah can fly
  - ...
- Now we ask: Hoopoe is a bird; can it fly?
- From your earlier observations
  - You create a generalized model of a bird
  - You extrapolate the properties to a new species of bird
- Induction is a special form of abduction

Wait till we study Hierarchial Bayesian Model


Case based reasoning



- Difference with Induction:
  - In induction, a generic model is formed
    - A new scenario is interpreted with the generic model
  - In CBR, no generic model is formed, cases exist in isolation
    - A new scenario is compared with earlier cases and the best match is used
- CBR can work with less experiential data

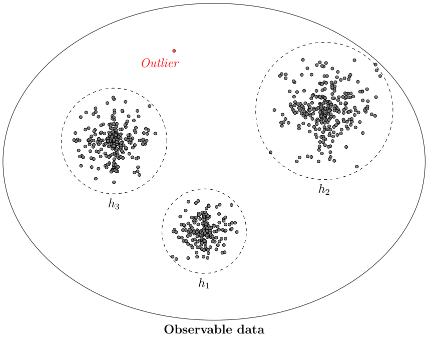
Exact match vs. Inexact match

- In real world, exact match is never possible
  - Inherent variations in the natural systems
  - Our knowledge about the world may be imprecise
- We resort to inexact match
  - Abduction: "Best" explanation
  - CBR: "Closest" case
- How to define the "best" or "closest"?
  - Objective measurement with "features"
  - The features are assumed to conform to metric space

Data driven reasoning

Machine learning

- Uses statistical similarity/associations to discover patterns
- We learn the models from data
- Flexible – no prior models
- Can't handle sparse and noisy data



Data driven reasoning

Example

	Sneeze	Cough	Body ache	Breathlessness
Patient 1	X	X	X	
Patient 2	X	X		
Patient 3		X	X	X
Patient 4		X	X	
Patient 5		X	X	X
Patient 6	X	X	X	
Patient 7		X		X
Patient 8	X		X	
Patient 9	X	X	X	
Patient 10			X	X

- No prior knowledge about diseases
- Patients 1, 2, 6, and 8 have similar symptoms $\rightarrow$ disease 1
- Patients 3,4,5,9 and 10 have similar symptoms $\rightarrow$ disease 2
- Patient 7 has Unique symptom
    - Observation error?
    - A new unknown disease?

- Pros: can discover new patterns (new models)
- Cons: inductive generalization not possible

Which one ?

- Which form of reasoning in used in the human mental processes ?
  - Probably all of them, depending on context
- Which form of reasoning in used in the human perception ?
  - Involves processing of sensory data (noisy)
  - Differences in visual appearance of object instances (uncertainties)
  - Incomplete model of the world (incomplete knowledge)
  - Abduction / Induction seem to be most appropriate

EdPuzzle: Bayesian Reasoning

Quiz

Quiz 03-01

End of Module 03-01