

Biological Vision and Applications

Module 01-01: About Biological Vision and the Course

Hiranmay Ghosh



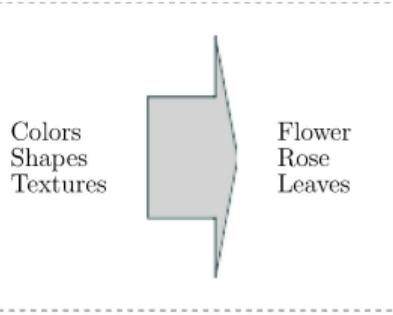
- Five sense organs to experience the world:
 - ▶ Eyes provide maximum information
- Vision is the process that transforms raw images to information
- This course is about study of principles of biological (human) vision
 - ▶ With an ulterior motive to apply them on computer vision systems

What does Human Vision System do?

Transforming visual signals to information



Image



Description

... This looks trivial !!
Let's have some insights

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



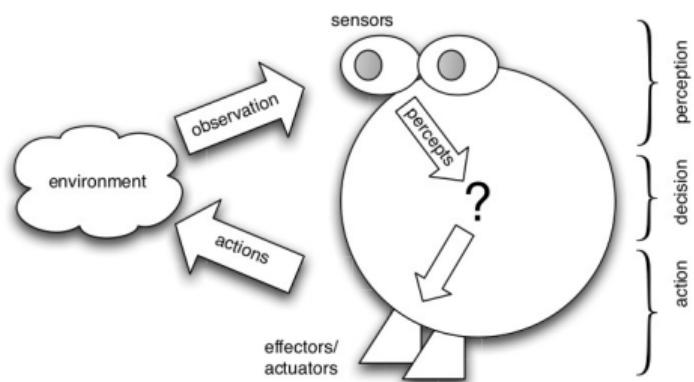
- **Identification**
 - ▶ Four players
 - ▶ Ball, Goalpost
 - ▶ Net, gallery, ...
- **Interpretation**
 - ▶ Football game
 - ▶ Free kick
- **Prediction**
 - ▶ Goal score?

- **Action**
 - ▶ Cheer (?)

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

Situated Vision

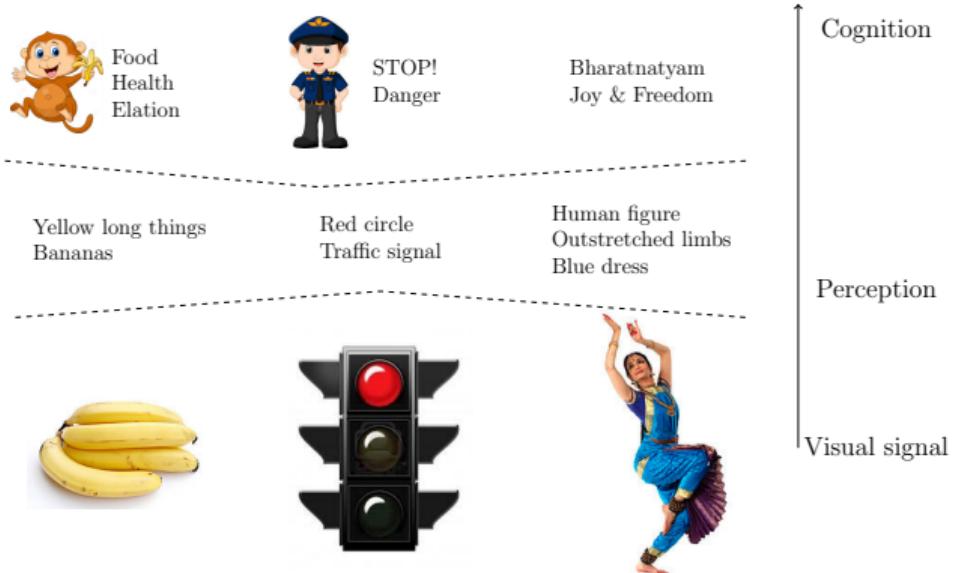
An eternal cycle of sensing and acting



- **Sensing:** Sense the environment
- **Processing:** Interpret the environment

- **Acting:** Influence the environment

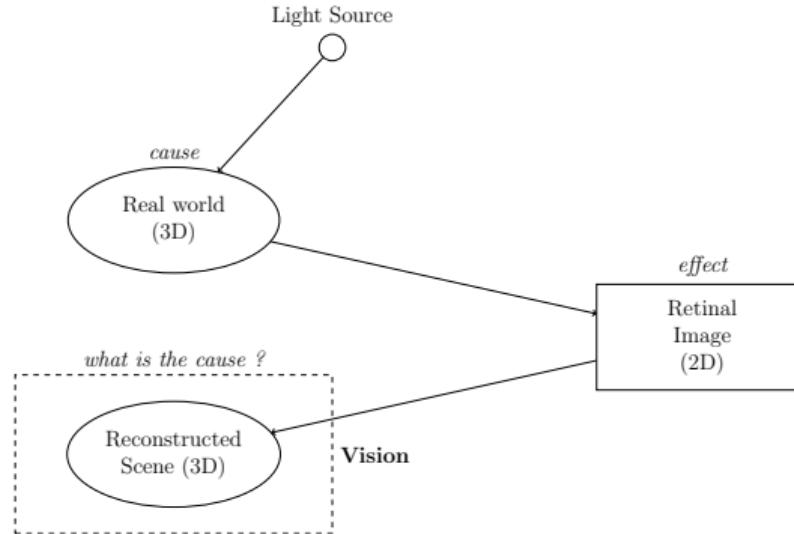
Layers of Interpretation



Goal of computer vision

- Emulate human vision system
 - ▶ ... Today's CV is far from achieving that
- How to bridge the gap?
 - ▶ Study principles of biological vision
 - ▶ Use them in computer vision algorithms

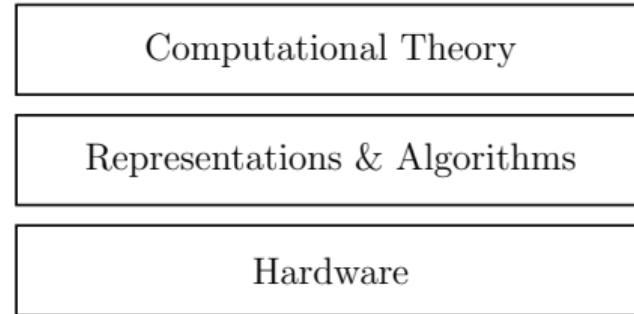
Computer Vision is an inverted problem



- Two approaches: Theory based and Machine learning based

Marr's abstractions (1976)

3-layer approach



- Independence of layers
- Same computational theories (as in biological vision system) can be implemented
 - ▶ On different hardware
 - ▶ With different representations and algorithms

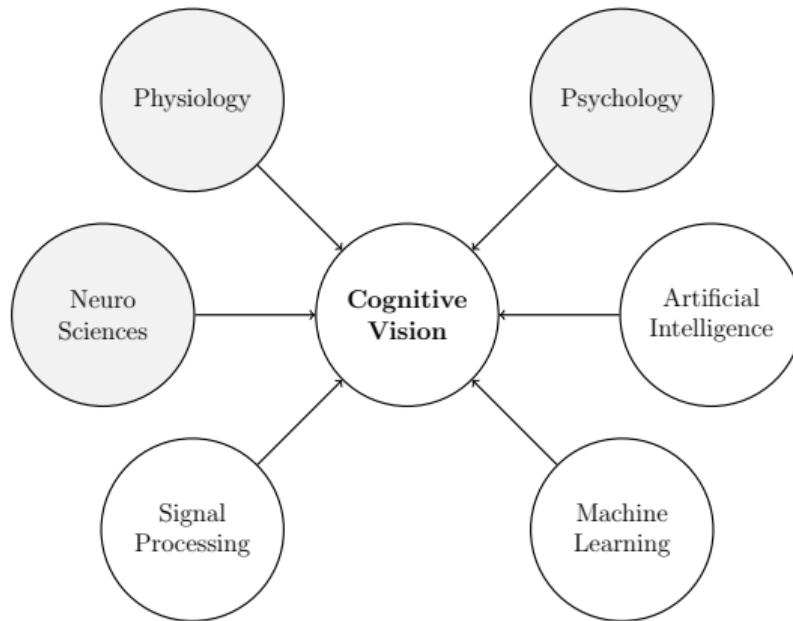
What are the challenges ?

- 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
 - ▶ Inter-class similarity: A red circular object may be a cricket ball or a tomato
 - ▶ Intra-class variations: No two roses are exactly identical!
 - ▶ Semantic gap
- Real-Time performance
 - ▶ Situated agents
 - ▶ Interacting continuously with the environment
 - ▶ Eternal cycle of sensing and acting

How does human mind cope up with the challenges ?

- Human mind does not have unlimited memory or processing power
 - ▶ Interpretation of a visual scene is mostly instantaneous and intuitively!
- How it happens ?
 - ▶ Data reduction: What to ignore and what to use?
 - ▶ Knowledge-based interpretation: Context, knowledge and experience
- Interpretation is subjective

What all are involved?



About the course

Pre-requisites

- **Mathematical/statistical skills**
- **Proficiency in programming**
 - ▶ Python: Colab env
 - ▶ Exposure to OpenCV library
- Computer Vision: Desirable, but not necessary
- AI / ML: Will be introduced as necessary
- Physiology / Psychology / Neurology: No

Study Material

- Textbook:
 - ▶ Hiranmay Ghosh. Computational Models for Cognitive Vision. Wiley-IEEE Press, 2020.
 - ▶ [Access link \(No download\)](#)
- Research papers & EdPuzzle videos will be announced in the class
- Lecture slides & videos will be uploaded in the classroom
- Background Study Material
 - ▶ [Metric Space and Image Features](#)
 - ▶ [Probability Theory](#)
- Recommend attending the class on a big screen (laptop / desktop)

Logistics

- Google Classroom code: y7b2wfi
- GMeet: [Link](#)
- Class Timings: **Slot R**
 - ▶ Thu: 6:00 - 7:30 pm
 - ▶ Sat: 4:00 - 5:00 pm
- **All course resources are accessible on IIT-J login only**

- **Continuous Evaluation [60]**

- ▶ Simple quiz at the end of every class (well almost!) [20]
 - ▶ Immediate deadline (**No second chance**)
- ▶ EdPuzzle assignments [15]
- ▶ Programming / non-programming assignments [20]
 - ▶ Programming assignment: Python (CoLab Environment)
 - ▶ **Submit readable code only** [[Example of unreadable code](#)]
- ▶ Class participation [5]
 - ▶ Attendance, Interactions, Timely submissions

- **Examinations [40]**

- ▶ Minor 1 & 2 [20]
- ▶ Major [20]

- **Plagiarism Policy: Zero Tolerance**

Quiz



No quiz for module 01-01

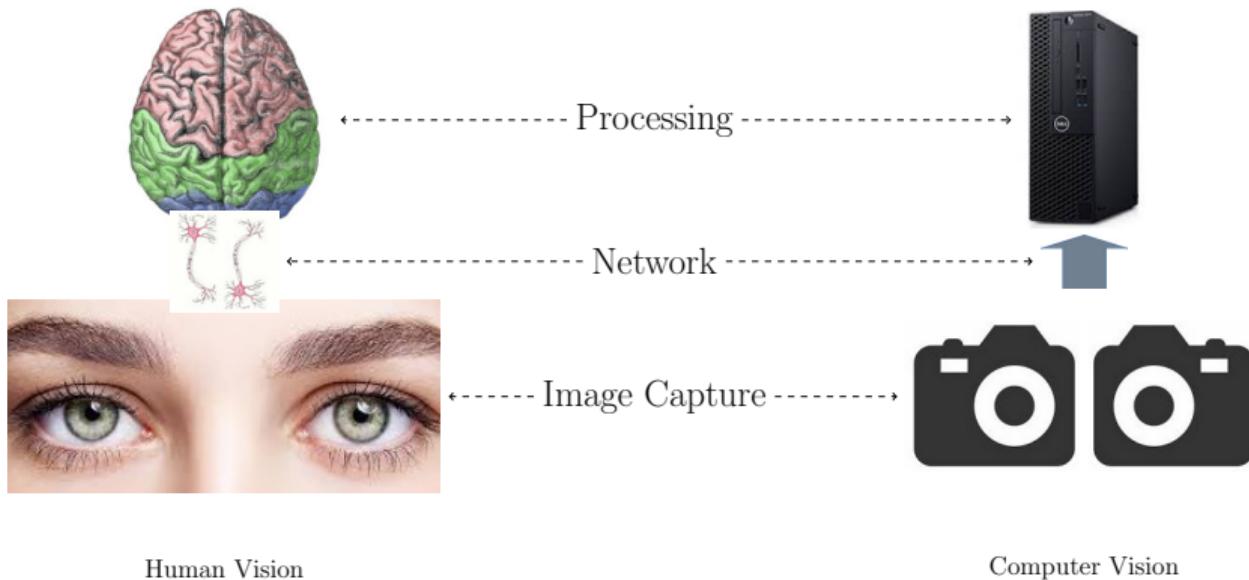
End of Module 01-01

Biological Vision and Applications

Module 01-02: Human Vision System

Hiranmay Ghosh

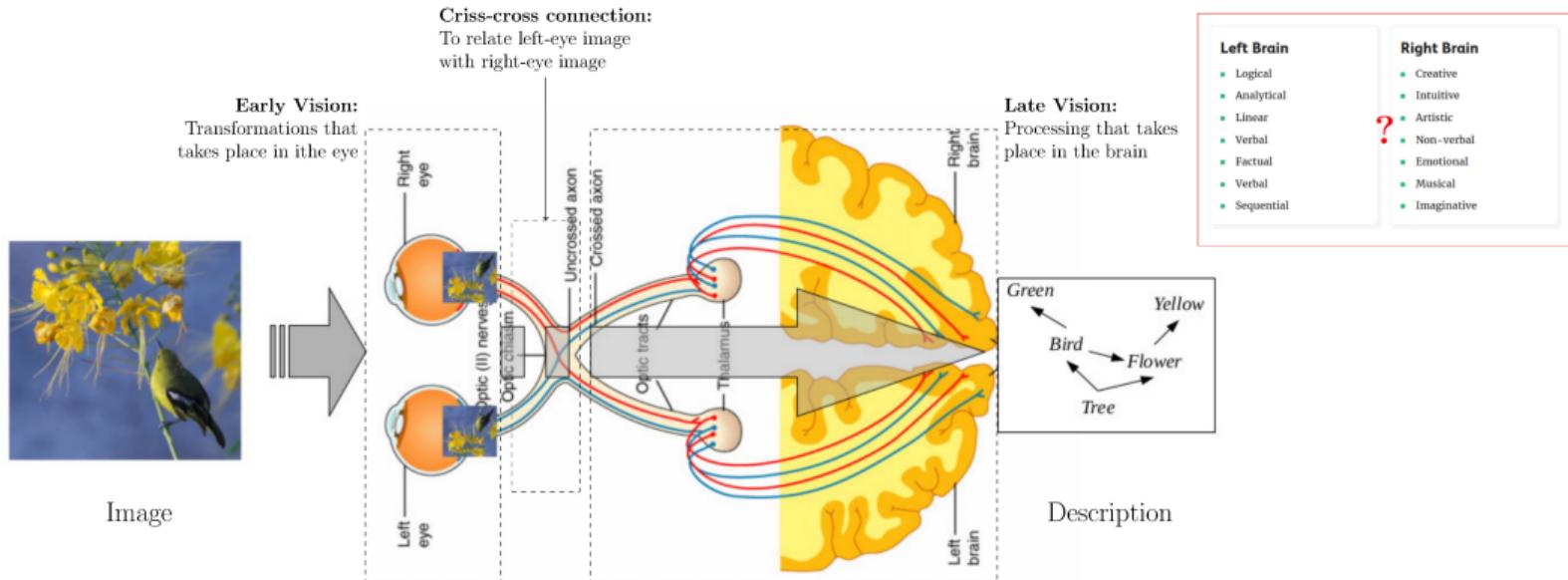
A Simplistic Analogy



... But, there is much more to it

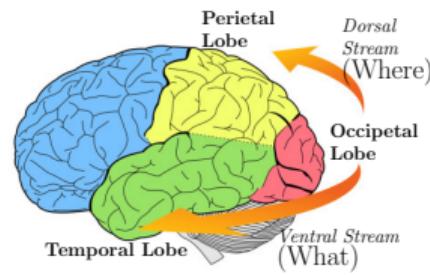
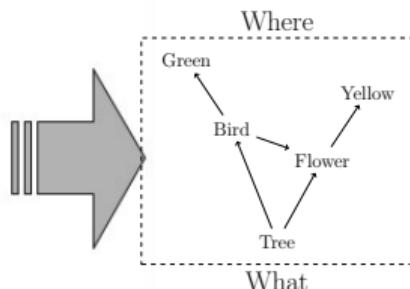
Overview of Human Vision System

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**

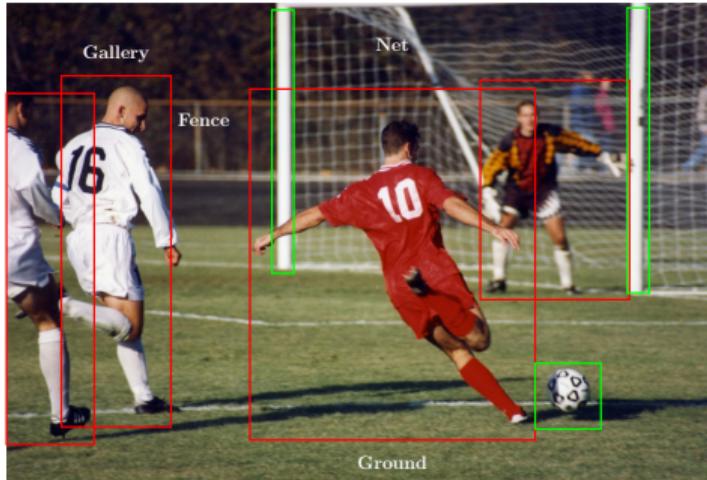
What is “Vision”?



- **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

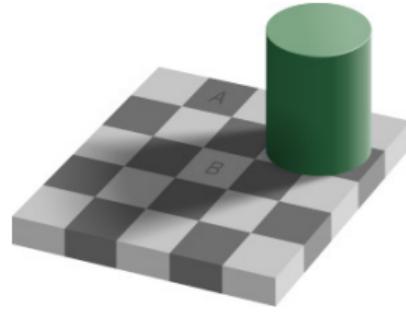


- Foreground and background
- **What ?**
 - ▶ Four Players, Ball, Goal Posts
 - ▶ Net, Fence, Gallery, Ground ...
- **Where ?**
 - ▶ Absolute & relative positions
 - ▶ Geometric organization

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”

Optical illusions



My Wife and My Mother-in-law (German postcard 1888)

Attention

Decide what is important



- We never look at the whole scene
 - ▶ 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

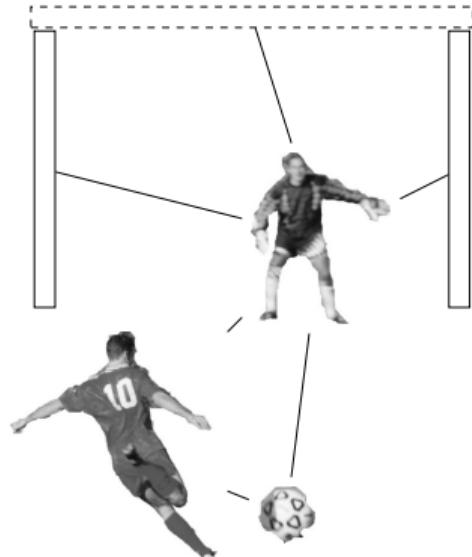
What is attention

- 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

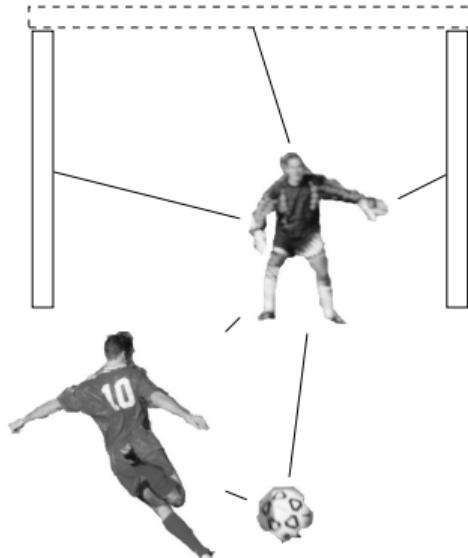
What do you “infer” from this picture ?



- 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
 - ▶ Information about this game, identity 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 ...

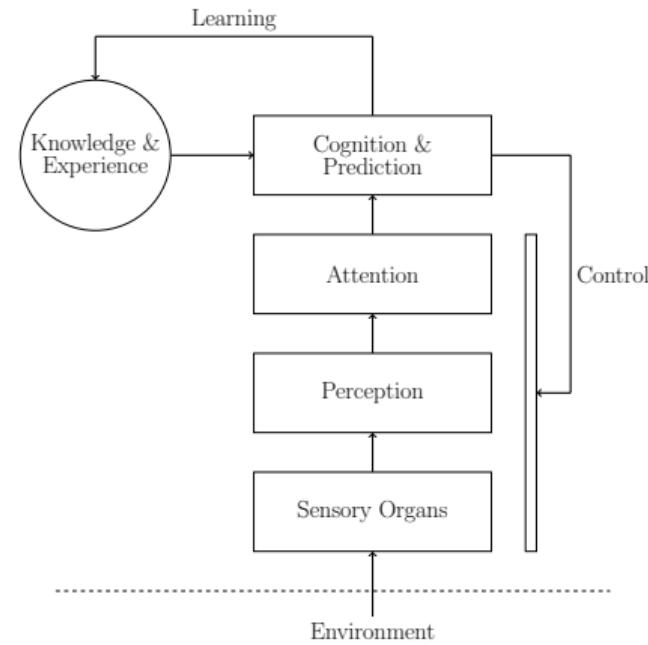
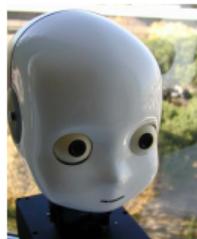
What is cognition

- 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

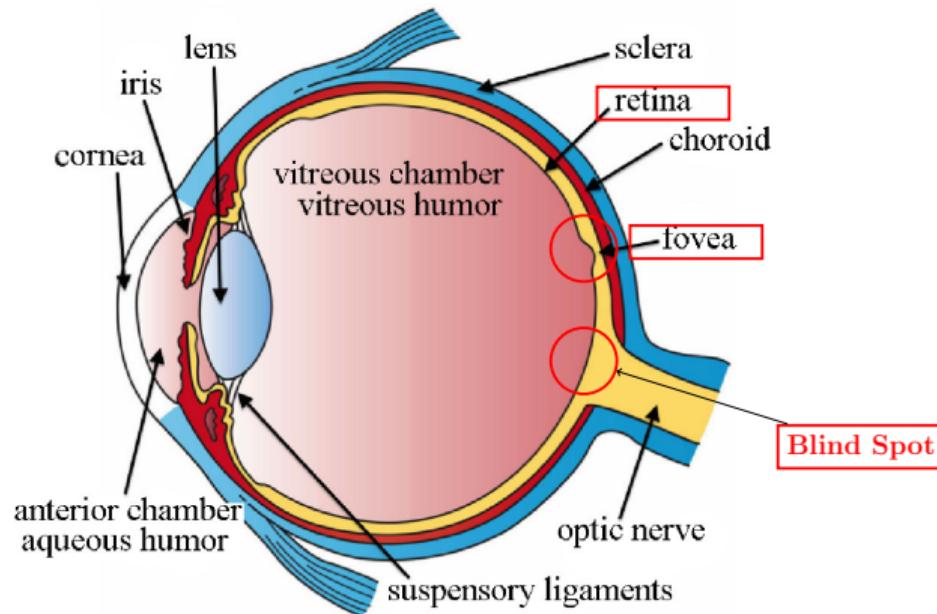
End of Module 01-02

Biological Vision and Applications

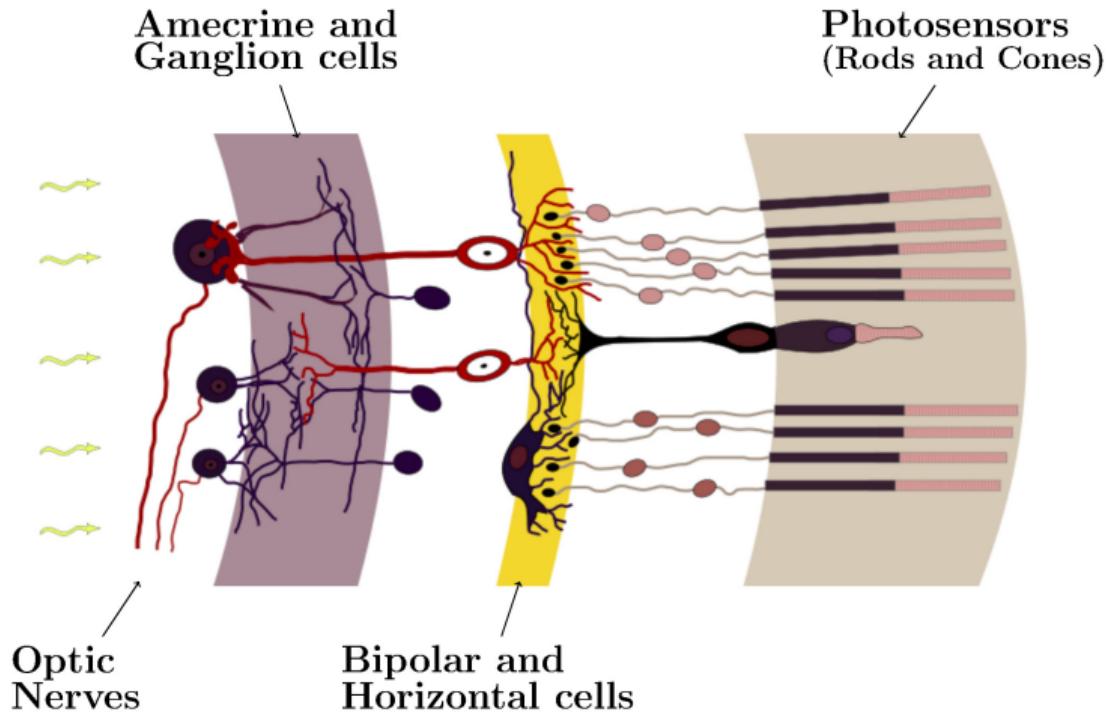
Module 02-01: Structure of the Eye

Hiranmay Ghosh

Structure of the eye



The Retina

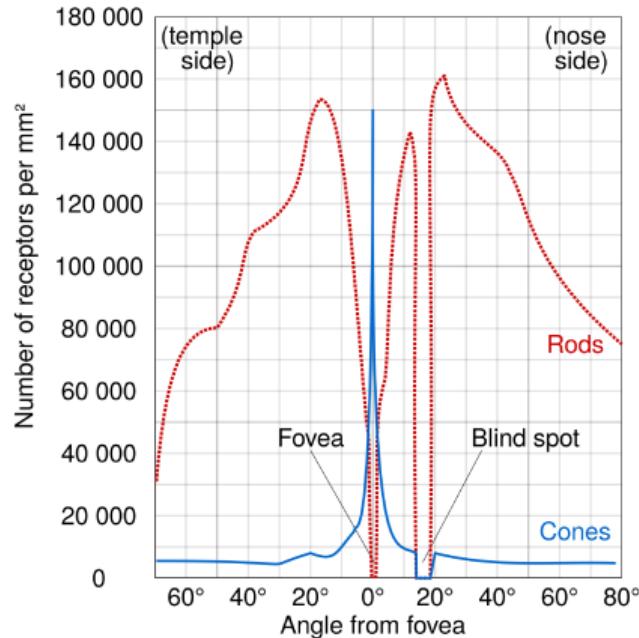


The Photosensors

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 fovea
 - ▶ Maximum acuity in this area
 - ▶ Best within 1.5 – 2 degrees
- Rods spread over 60 – 80% of visual field
 - ▶ This area is called the peripheral area
 - ▶ Resolution decreases linearly with the distance from the center of the eye

Attention, Fixation and Saccade

The diagram shows three lines of black text with red arrows indicating eye movement paths. The top line reads "Mark had a new bike. The bike was red. One day". The middle line reads "Mark rode his bike to the park. Mark left his new bike". The bottom line reads "by a tree. Mark played on the slide. He played on the". Red arrows above the text show a series of saccades (fast eye movements) between words, with small curved arrows indicating the direction of gaze during each fixation (the brief pause between saccades).

Mark had a new bike. The bike was red. One day

Mark rode his bike to the park. Mark left his new bike

by a tree. Mark played on the slide. He played on the

- 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

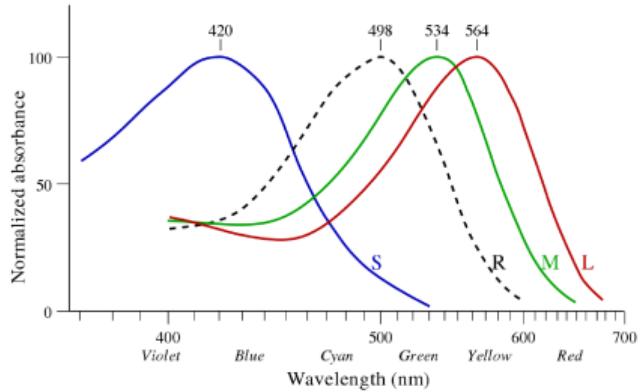
The uneven distribution of the cones on the retina results in high acuity image to be formed only in the foveal region, which covers about $1.5 - 2^\circ$ of the visual field. Best acuity occurs at the *fovea centralis* that is about $\frac{1}{10}$ th of the fovea. When a person looks at an object, it 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

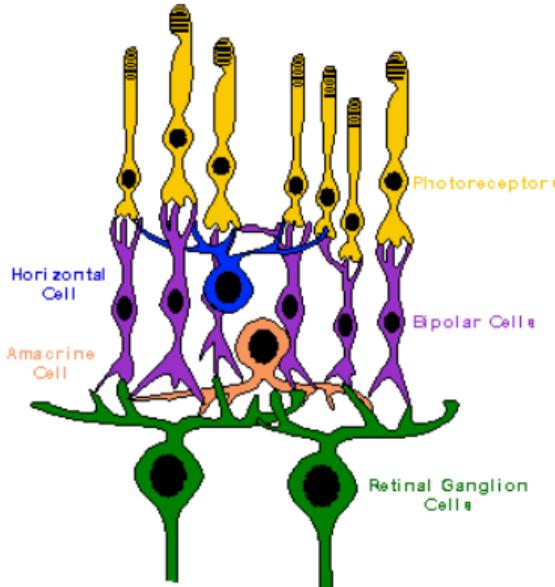
The uneven distribution of the cones on the retina results in high acuity image to be formed only in the foveal region, which covers about $1.5 - 2^\circ$ of the visual field. Best acuity occurs at the *fovea centralis* that is about $\frac{1}{10}$ th of the fovea. When a person looks at an object, it 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
- Amacrine cells
 - ▶ Interconnects the ganglions

We shall look at their functions in the following modules

Quiz



Quiz 02-01

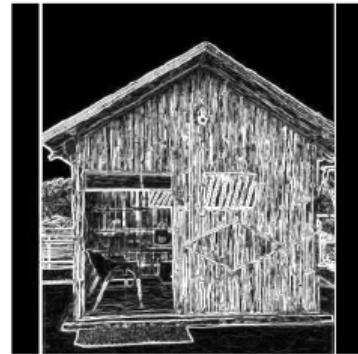
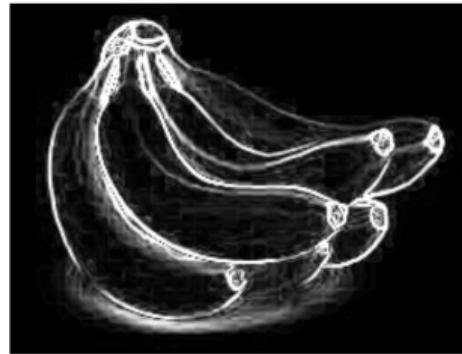
End of Module 02-01

Biological Vision and Applications

Module 02-02: Edge detection

Hiranmay Ghosh

Why edge detection is important

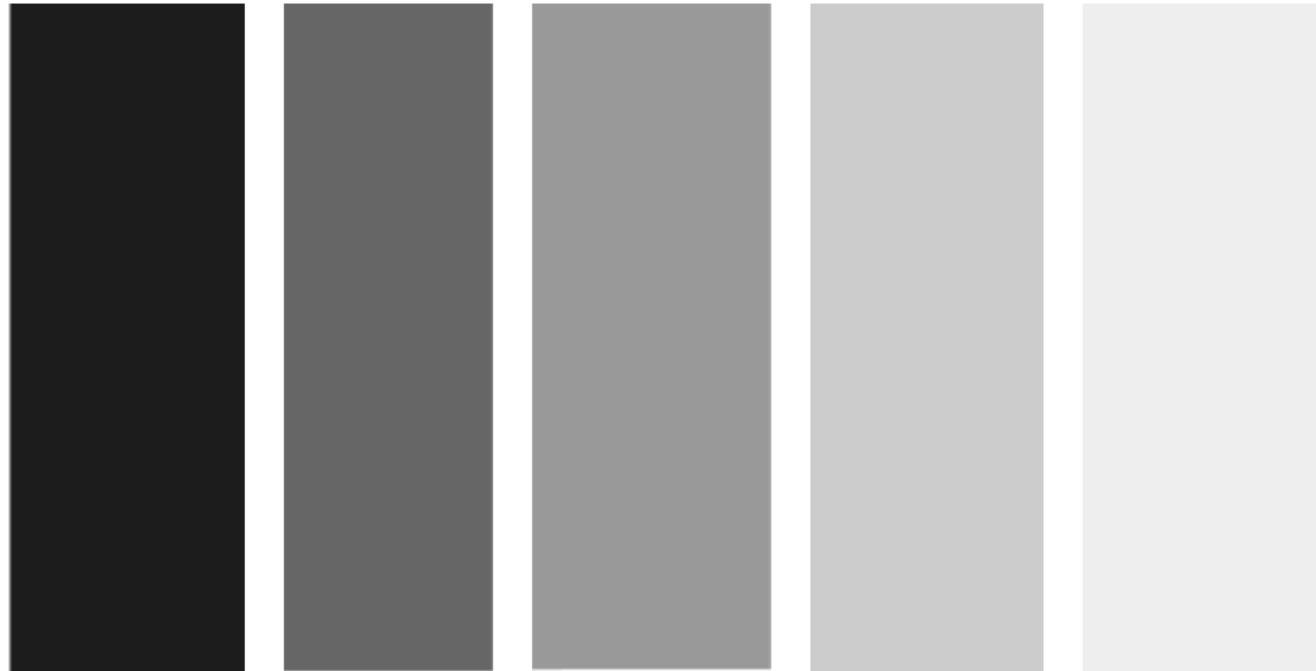


- Most of the objects are identified by their shapes

- How are the shapes detected ?

Machband Effect

Bars of uniform illumination

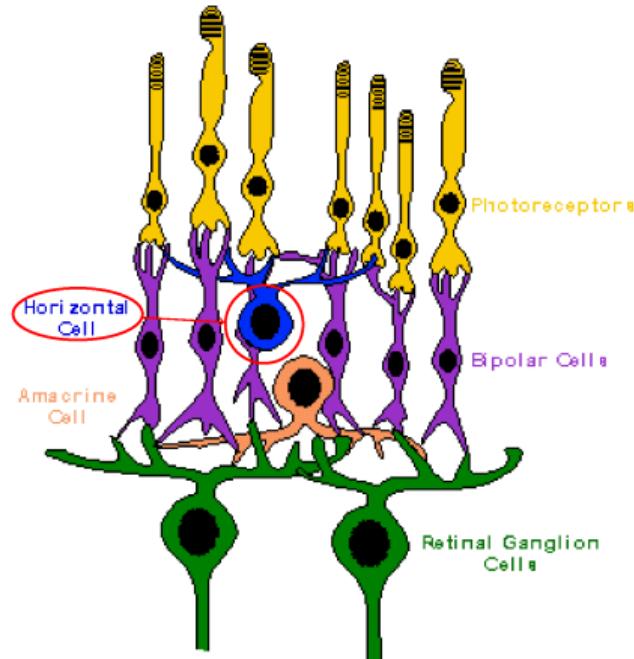


Machband Effect

Adds contrast to vision

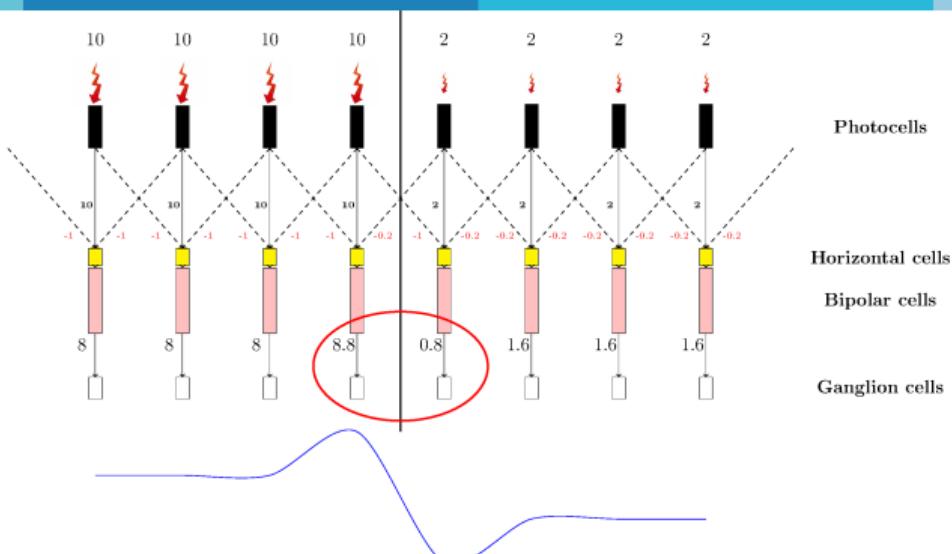


The Horizontal cells



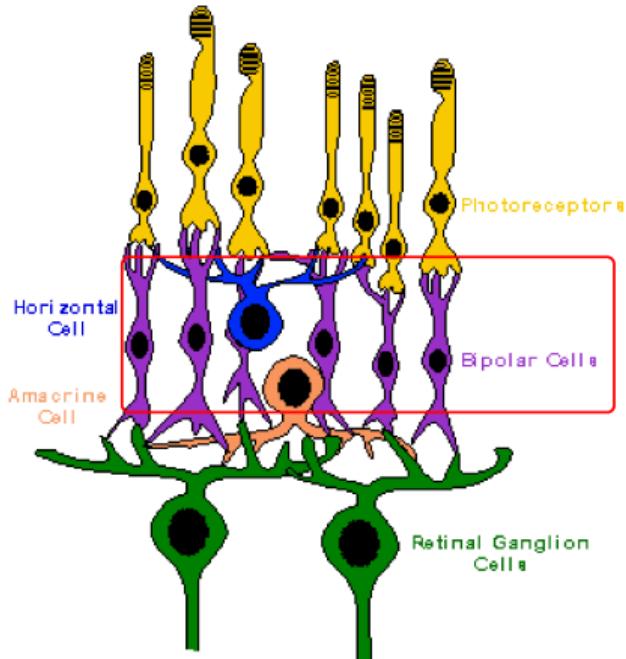
- Connects nearby photo-receptors of the same kind
- Carries a **Lateral Inhibition** signal
 - ▶ proportional to the response level

Computational Model



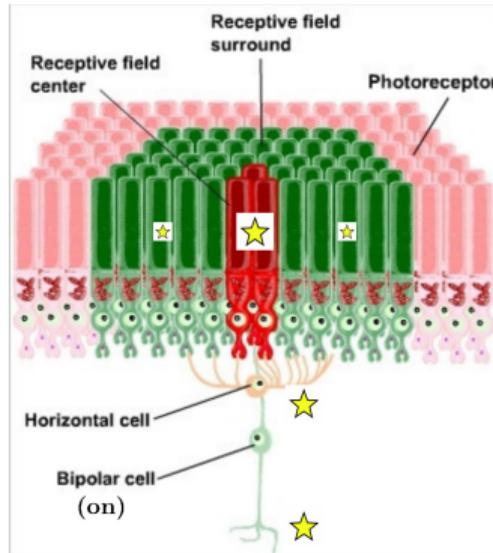
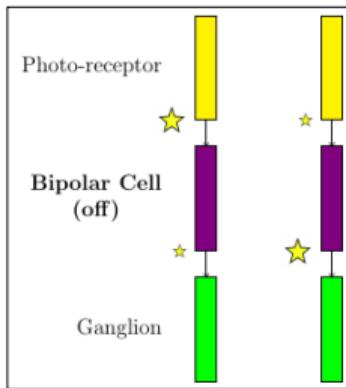
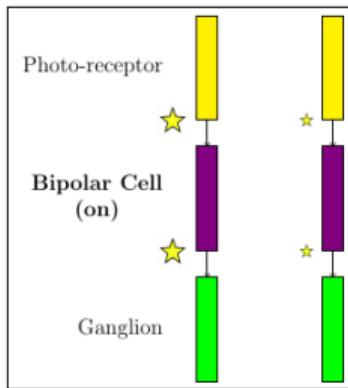
This example assumes $\frac{1}{10}$ th of response to be transmitted as inhibition signal

The Bipolar cells



- Photo-receptors are connected to Ganglions through bipolar cells
- 10 million bipolar cells connects to 125 million photo-receptors
 - ▶ Connects to few cones (even 1)
 - ▶ Connects to many rods
- Data reduction

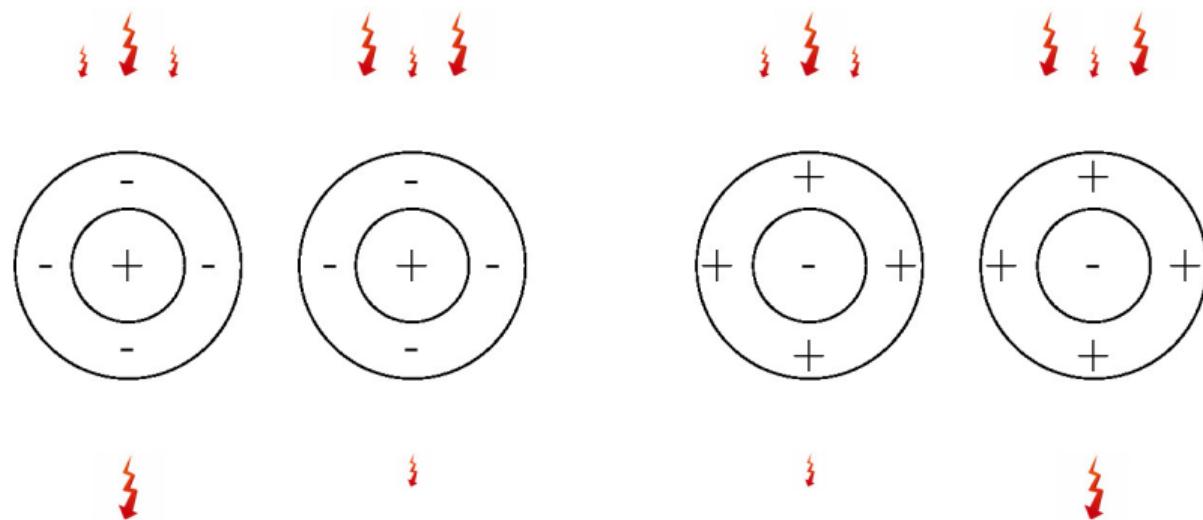
Center-Surround effect



The ganglion output is a weighted linear combination of excitation levels of the photoreceptors in the receptive field.

Center-Surround effect

... continued

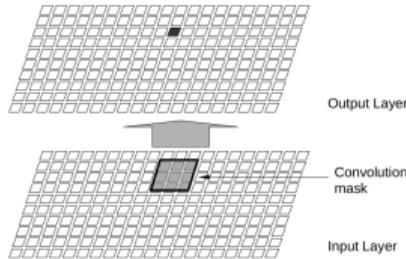


On-center configuration
“On” bipolar cell at center

Off-center configuration
“Off” bipolar cell at center

Generic model of signal processing in early vision

Digital Convolution (2D)



- Compute weighted average of surrounding pixels
- No data reduction at this stage

Inputs:

Image $I = [[I_{xy}]]$, $x = 1 : W, y = 1 : H$

Mask $M = [[M_{xy}]]$, $x, y = -m : +m \quad (m << W, H)$

Output:

Transformed Image $I' = [[I'_{xy}]]$, $x = 1 : W, y = 1 : H$

$$\text{where } I'_{x,y} = \sum_{i=-m}^{+m} \sum_{j=-m}^{+m} M_{x+i,y+j} \cdot I_{x-i,y-j}$$

Fundamental operation in a Convolutional Neural Network (CNN)

Image filters

$\frac{1}{273}$

1	4	7	4	1
4	16	26	16	4
7	26	41	26	7
4	16	26	16	4
1	4	7	4	1

Laplacian filters

-1	-1	-1
-1	8	-1
-1	-1	-1

on-center

1	1	1
1	-8	1
1	1	1

off-center

- Integrating filters

- ▶ Gaussian filter (with $\sigma = 1$)
- ▶ Used for noise reduction

- Differentiating filters

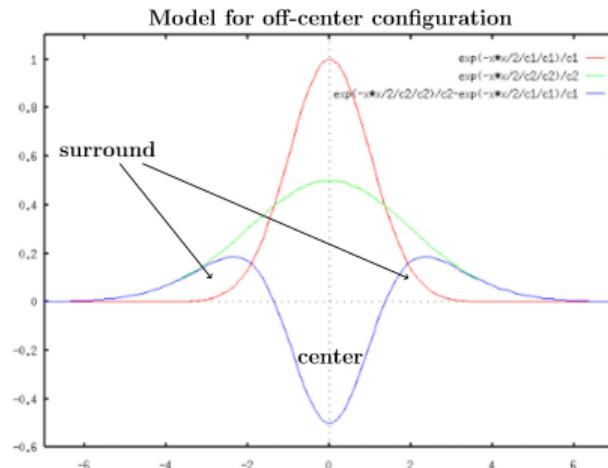
- ▶ Implements center-surround operation
- ▶ Used for contrast enhancement

In classical image processing, the filters used to be hand-crafted

In CNN, they are machine learnt

Difference of Gaussian

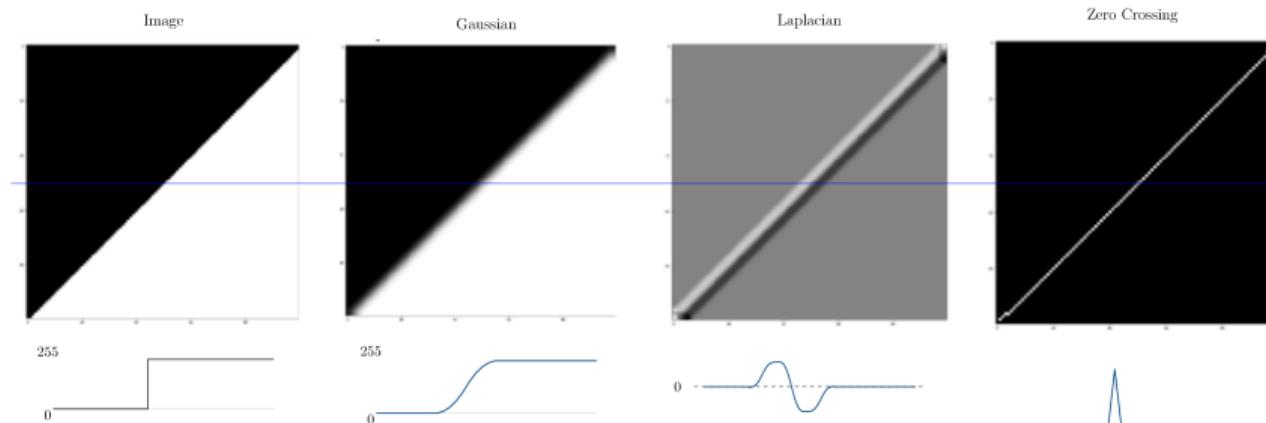
Approximates Laplacian of Gaussian (LoG)



- Gaussian smoothing (noise reduction) precedes differentiation
- DoG is an approximation of LoG

Edge detection

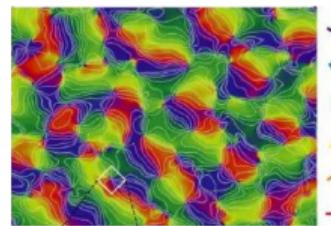
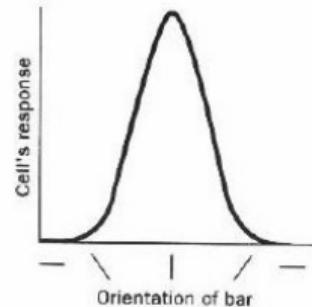
Zero-Crossing of LoG/DoG



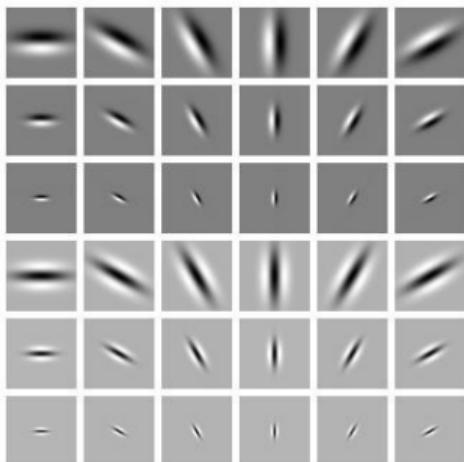
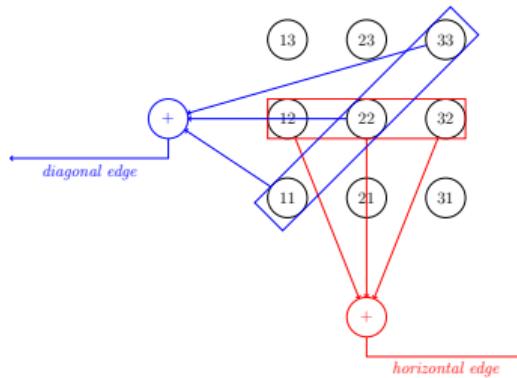
- Orientation ?

Detection of Edge Orientations

- It has been observed that
 1. Some cells in the visual cortex respond to edge orientations
 2. Different cells respond to different orientations
 3. The cells that responds to same orientation are grouped together



Oriented Filter Banks



- Ganglions connect to orientation detection cells in different combinations
- Results in edges with different orientations to be detected
- Motivates design of Gabor filters

[More in EdPuzzle](#)

Convolution filters for edge detection

Sobel filters

+1	+2	+1
0	0	0
-1	-2	-1

-1	0	+1
-2	0	+2
-1	0	+1

- Integration and Differentiation

$$\mathbf{G}_x = [1 \ 2 \ 1]^T * ([+1 \ 0 \ -1] * \mathbf{I})$$

$$\mathbf{G}_y = [+1 \ 0 \ -1]^T * ([1 \ 2 \ 1] * \mathbf{I})$$

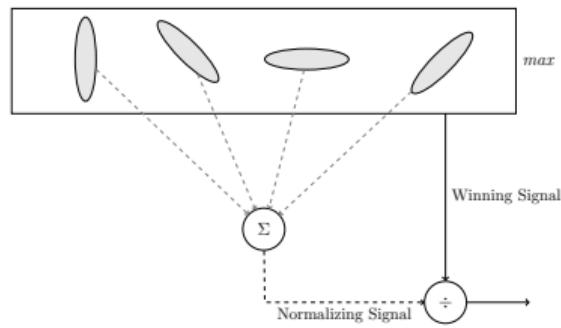
$$|\mathbf{G}| = \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2}$$

$$\theta = \tan^{-1} \frac{\mathbf{G}_y}{\mathbf{G}_x}$$

Does not happen this way in brain

Winner Take All (WTA) and Normalization

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
 - ▶ The winner needs to stand out to produce strong response

Transformation in the eye



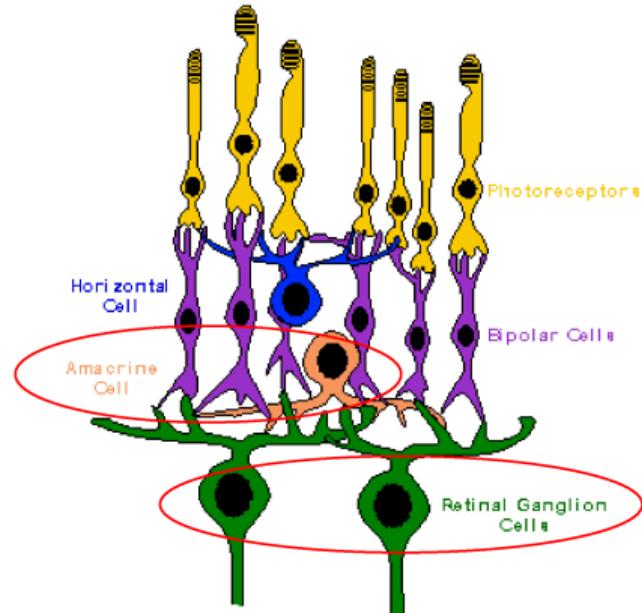
Retinal image



Neural image

- Human vision is sensitive to contrast, and not to brightness
- There is huge data reduction
 - ▶ 126 mn photo-receptor → 10 mn bipolar cells → 1 mn optic nerves (approx)

The amacrine and the ganglion cells



- Ganglions connect to the brain
- Amacrine cells contribute to motion detection

Quiz



Quiz 02-02

End of Module 02-02

Biological Vision and Applications

Module 02-03: Motion perception

Hiranmay Ghosh

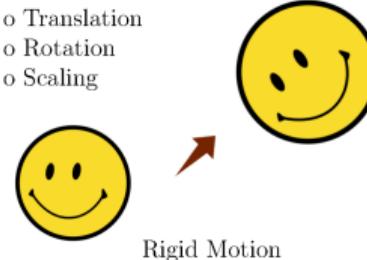
Why motion detection is important

- Distinguish objects of interest (foreground) from the background
- Determine change of location (where) of the objects of interest with time



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



Elastic Motion

- We shall mostly talk about rigid motion

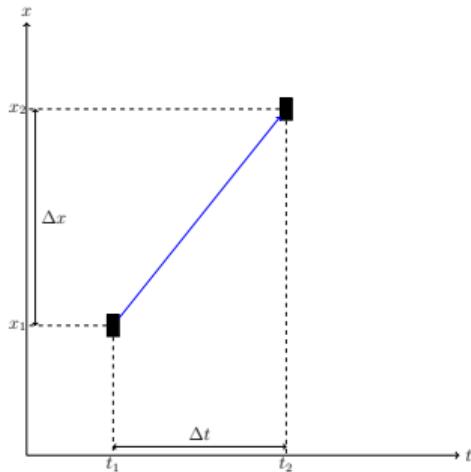
Continuous and Discrete Motion

- Human observers can distinguish two types of motion: **Continuous** and **Discrete**
 - ▶ For **perception** of continuous motion, an object need not move continuously over the retinal field
 - ▶ **Examples:** Alternately blinking festive lights, movie / TV
- There are two stages of motion detection
 - ▶ **Short range** (60 - 100ms, 10 - 15' of visual arc): Based on local intensity changes
 - ▶ Local contrasts: early vision
 - ▶ Continuous motion
 - ▶ **Long range** (400ms): Based on token matching
 - ▶ Object recognition: late vision
 - ▶ Discrete motion
- We shall talk about short range motion detection first

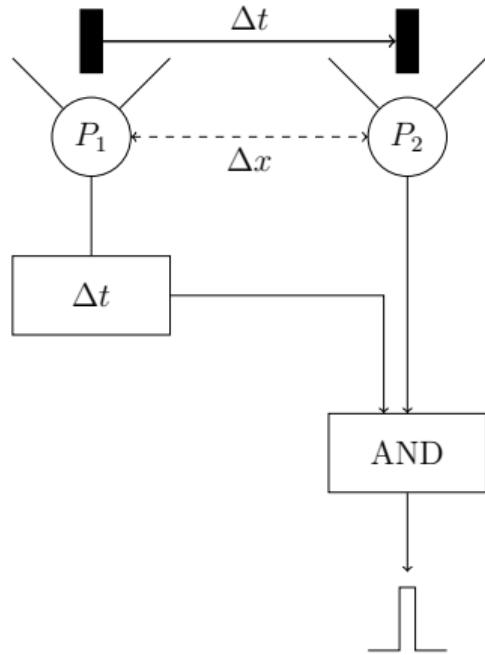
Motion is a correspondence problem

- An image in motion: $I(x, y, t)$
- The motion: $\vec{v}(x, y, t)$ – How to estimate from values of $I(x, y, t)$ over t ?
 - ▶ Sometimes it is sufficient to detect motion

- Detecting motion
 - ▶ Same object appears at (x_1, t_1) and at (x_2, t_2)
 - ▶ If for some (t_1, t_2) , $(x_1 \neq x_2)$
- Measuring motion
 - ▶ Velocity $v = \frac{\Delta x}{\Delta t}$



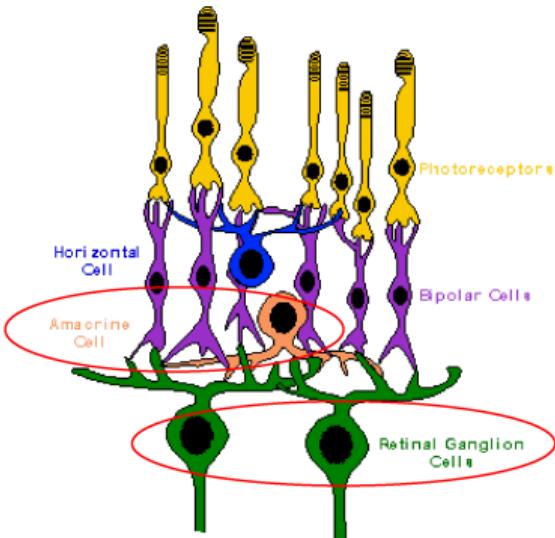
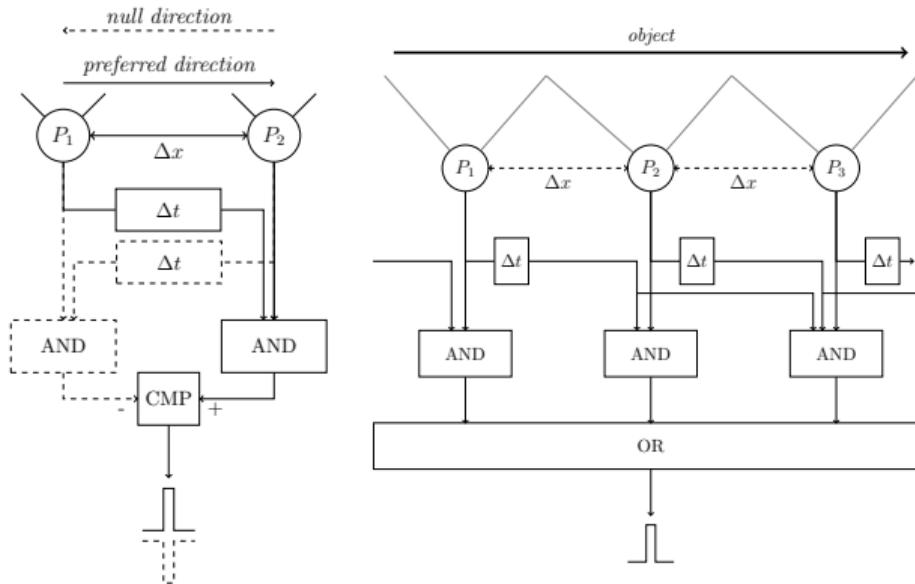
Reichardt model for motion detection



- Reichardt model
 - ▶ Eyes of house-flies
- Obvious limitations
 - ▶ Motion can be detected in one direction only
 - ▶ Motion can be detected only when $v \approx \frac{\Delta x}{\Delta t}$
 - ▶ Noise may induce false positives / negatives

Reichardt model for motion detection

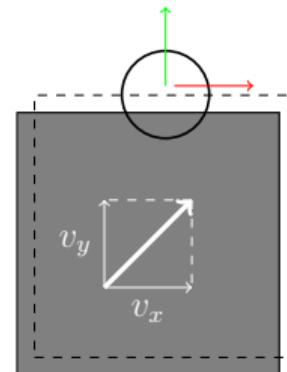
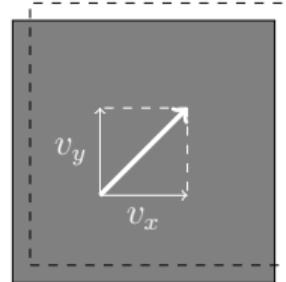
Addressing the limitations



Intensity based scheme

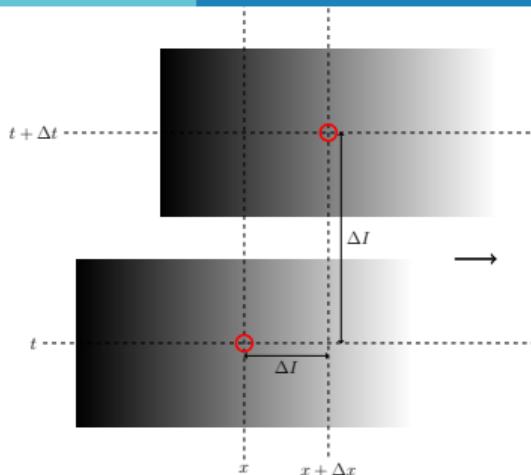
Estimating local motion

- See the contour changing and infer motion
 - ▶ Works when there is a significant intensity (or color) variation



Aperture problem: Motion can be perceived in the direction of intensity change only

Gradient model of motion estimation



$$I(x, t) = I(x + \Delta x, t + \Delta t)$$

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

$$\Delta I = I(x + \Delta x, t) - I(x + \Delta x, t + \Delta t)$$

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

$$v_x = \frac{\Delta x}{\Delta t} = -\frac{\partial I / \partial t}{\partial I / \partial x}$$

$$v(z, t)_{\nabla I} = -\frac{I_t(z, t)}{\nabla I(z, t)}$$

where

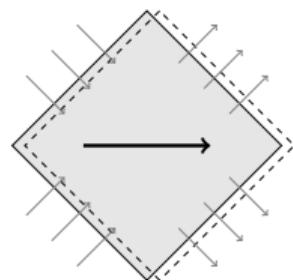
$v(z, t)_{\nabla I}$ = the local velocity at z , in the direction of the spatial intensity gradient

$I_t(z, t)$ = the temporal gradient for local illumination change

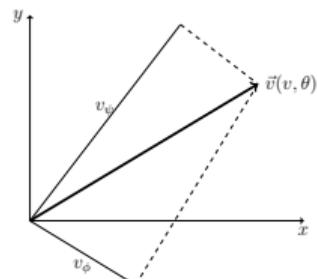
$\nabla I(z, t)$ = the spatial gradient for local illumination change

Rigid motion in image plane

Constant velocity assumption (translation only)



- The overall 2D motion can be estimated from the perceived motion at various points on the contour.
- Use many points for error-resilience



Given $v_\phi, v_\psi, v = ?, \theta = ?$

$$v_\phi = v \cdot \cos(\theta - \phi)$$

$$v_\psi = v \cdot \cos(\theta - \psi)$$

Solve for v and θ

$$v \cdot \cos(\theta - \phi_1) = v_1$$

$$v \cdot \cos(\theta - \phi_2) = v_2$$

$$v \cdot \cos(\theta - \phi_3) = v_3$$

...

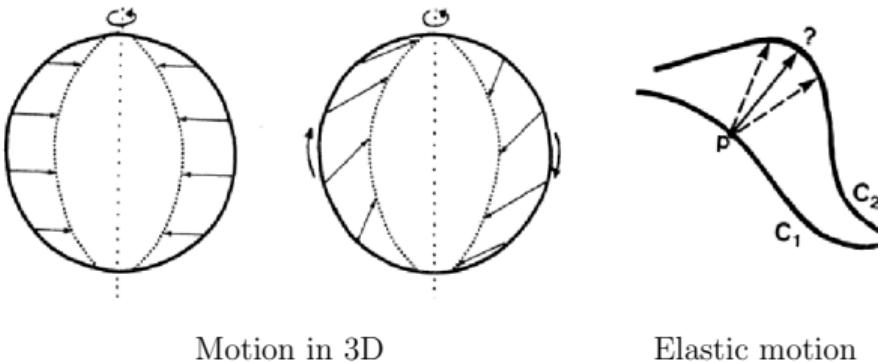
...

$$v \cdot \cos(\theta - \phi_n) = v_n$$

- SVD through example

Ambiguity in motion estimation

More general cases



- Sources of information loss
 - ▶ Projection of 3D object to 2D image
 - ▶ Projection of movement to intensity variation
- $\vec{V} = v_{\perp} \cdot \vec{u}_{\perp} + v_{\top} \cdot \vec{u}_{\top}$ (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 02-03

End of Module 02-03

Biological Vision and Applications

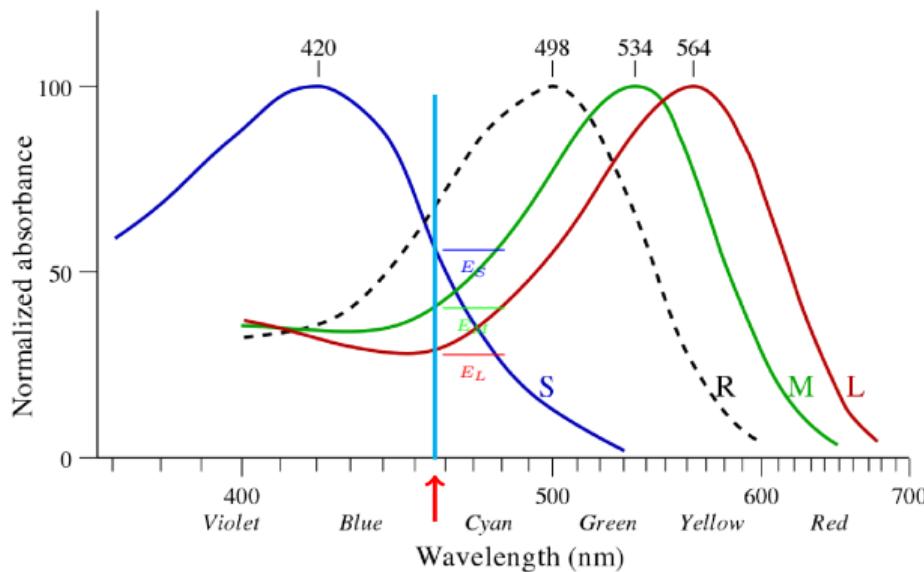
Module 02-04: Color Perception



Hiranmay Ghosh

Color perception

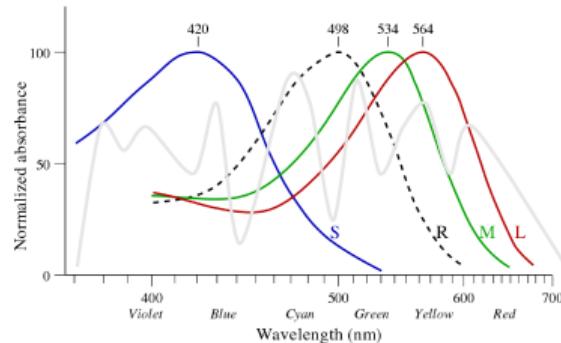
... not a property of spectral component of light, but how your eyes respond to it



- Perceived color $C = f(E_S, E_M, E_L)$

Color perception

Color is an emergent entity, distinct from properties of light



- Incident light: $I(\lambda)$
- Excitation levels of the cones are given by

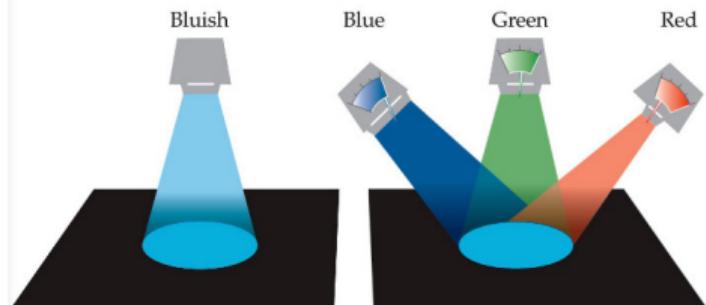
$$\begin{aligned}E_S &= \int_{\lambda} S(\lambda).I(\lambda).d\lambda \\E_M &= \int_{\lambda} M(\lambda).I(\lambda).d\lambda \\E_L &= \int_{\lambda} L(\lambda).I(\lambda).d\lambda\end{aligned}$$

- Perceived color $C = f(E_S, E_M, E_L)$
- **Metamers:** Lights with different spectral components that give rise to same color perception

EdPuzzle: Is your red the same as my red ?

Trichromatic Color Theory

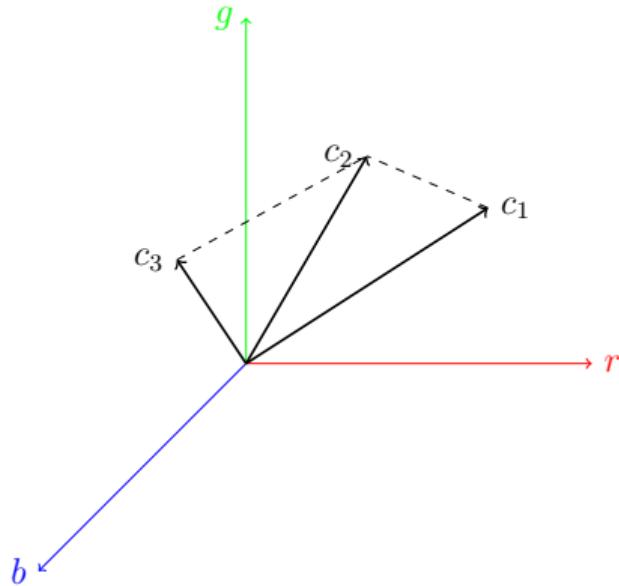
- A perceived color can be matched by a combination of three **primary colors**
 - ▶ Proved by psychological experiments
 - ▶ By convention, R, G and B are taken as primary colors



The three colors need not necessarily be “Blue”, “Green” and “Red”

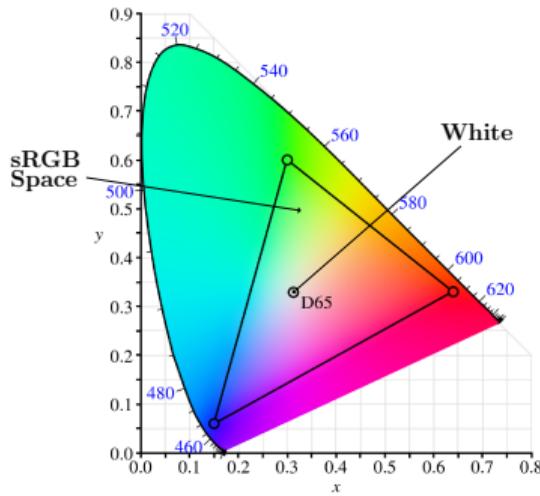
Device dependent color models

RGB Model



- Electronic devices typically use “red”, “green” and “blue” color guns
- Each color is represented by a point in 3D space
 - ▶ $\vec{c} = \{\alpha.\vec{r} + \beta.\vec{g} + \gamma.\vec{b}\}$
 - ▶ Are the vectors $\vec{r}, \vec{g}, \vec{b}$ orthogonal?
- Let \vec{c}_1, \vec{c}_2 and \vec{c}_3 represent three colors in rgb space
 - ▶ $|\vec{c}_1 - \vec{c}_2| < |\vec{c}_2 - \vec{c}_3|$ does not necessarily mean that
 - ▶ \vec{c}_2 is perceptually closer to \vec{c}_1 than \vec{c}_3

sRGB Color space

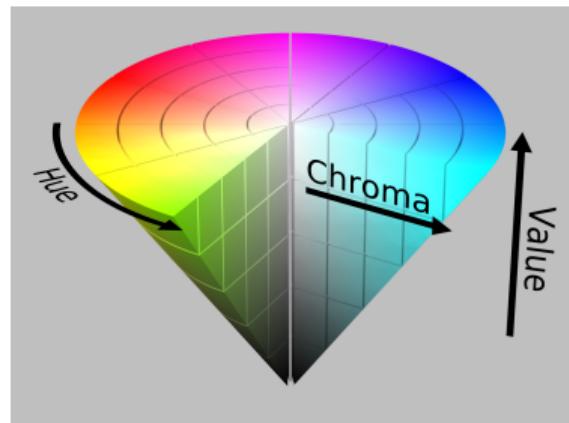


- Electronic devices use three primary color guns
- Perceived color can be matched by a “linear combination” of three primary colors
- Unfortunately, we can only add (not subtract) color in electronic devices
 - ▶ We can produce only a subset of perceivable colors with the devices
- The color space that can be produced by a device is called sRGB space
 - ▶ Depends of the device characteristics

Device independent color models

HSV Model, CIE Model

- Munsell described color in terms of its three perceptual properties, namely
 - ▶ Hue (shade), Value (brightness), and Chroma (color purity)
- ... 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



Merits of Trichromatic Color Theory

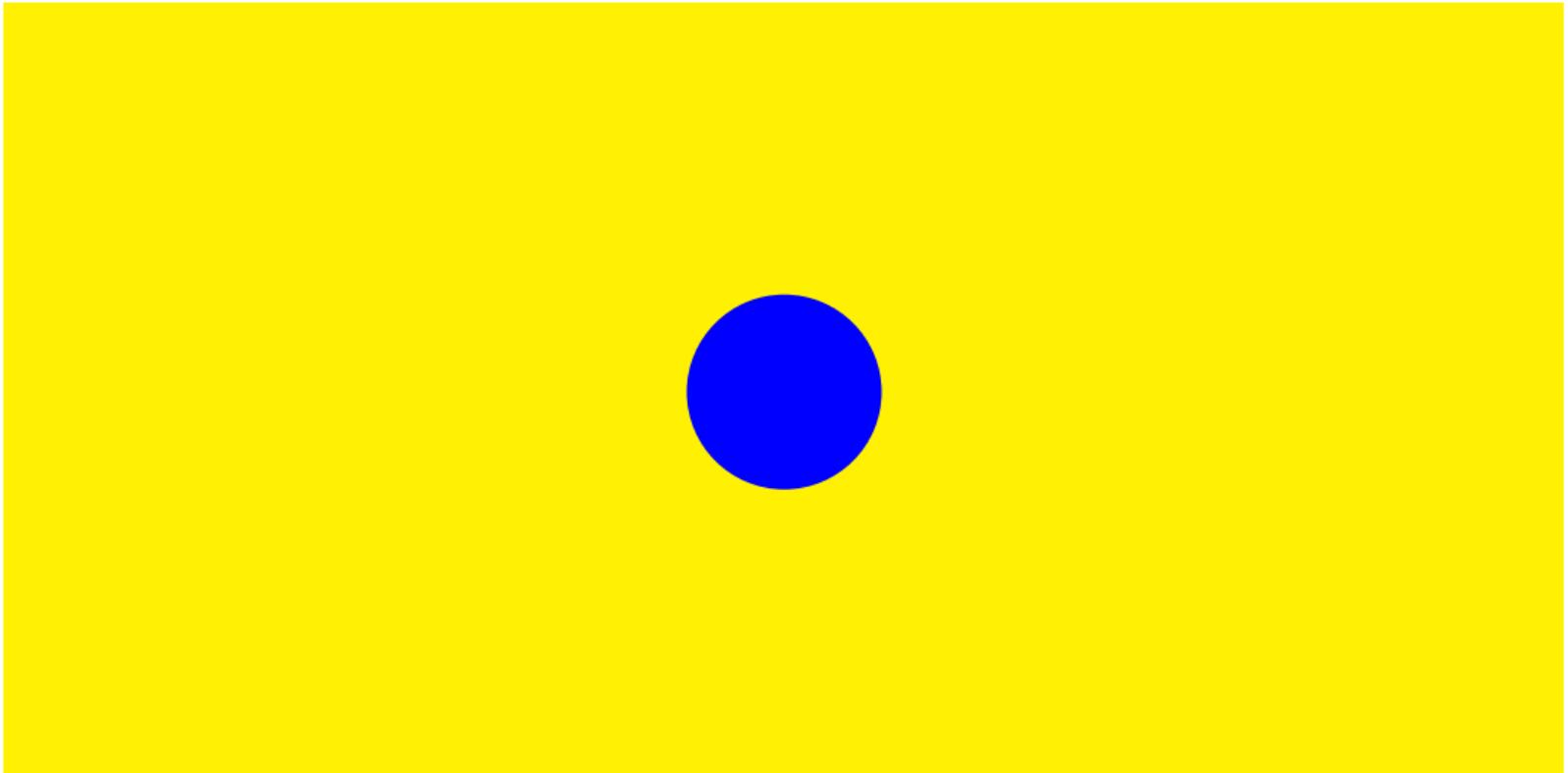
- Accounts for the 3 independent dimensions of color space
- Can explain metameristic matching
 - ▶ Mixing of 3 “primary” colors is sufficient to match any other visible color.
- Partially accounts for color blindness

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 color shades
- 24 bits has a provision to represent 16 million color shades!
 - ▶ The issue is how we intelligently utilize the 24 bits

After-images (experiment)

Concentrate on the picture below for about 15 sec



After-images (experiment)

... Contd.

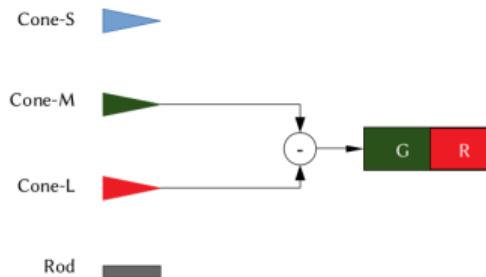
Opponent (color) process theory

- Opponent colors show up as after-image
 - ▶ Because of “fatigue” of the photoreceptors
- 3 pairs of opponent colors (observed through psychological experiments)
 - ▶ red vs. green
 - ▶ blue vs. yellow
 - ▶ dark (black) vs. bright (white)

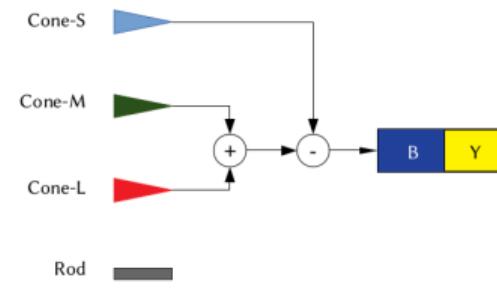
Opponent process theory

(Continued)

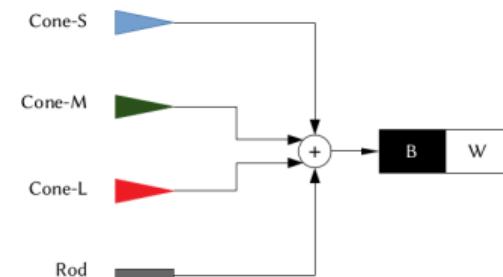
- Opponent color contrasts are explained by neural connections
- 3 new (derived) color channels are formed



$$GR = E_L - E_M$$



$$BY = E_S - (E_L + E_M)$$

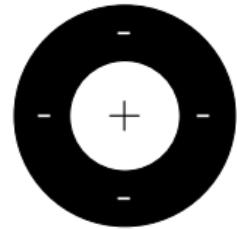
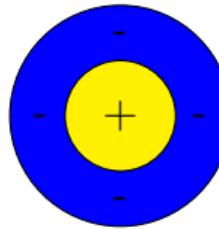
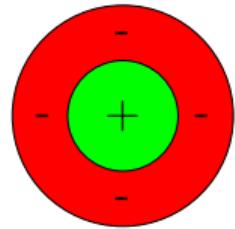
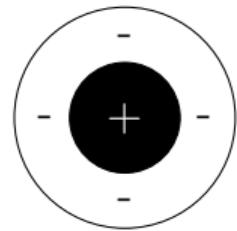
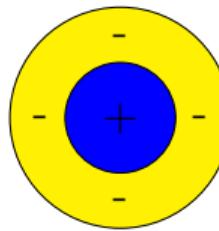
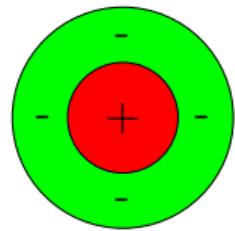


$$BW = E_L + E_M + E_S + E_R$$

Opponent process theory

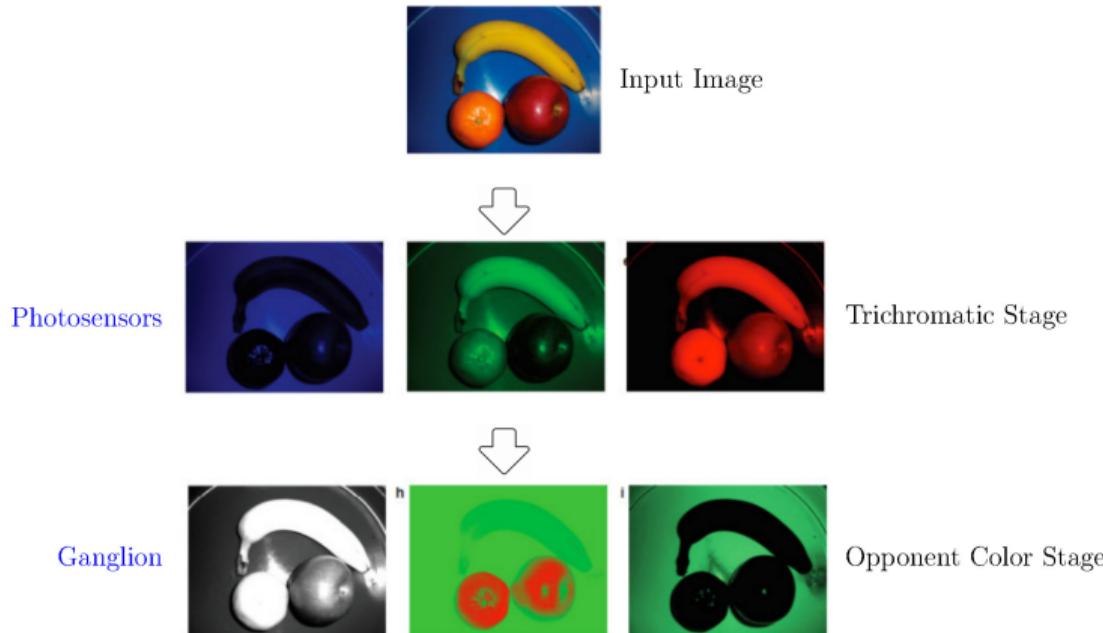
(Continued)

- Different organizations for opponent color sensitive cells



Dual (color) process theory

Stages of color processing



Quiz



Quiz 02-04

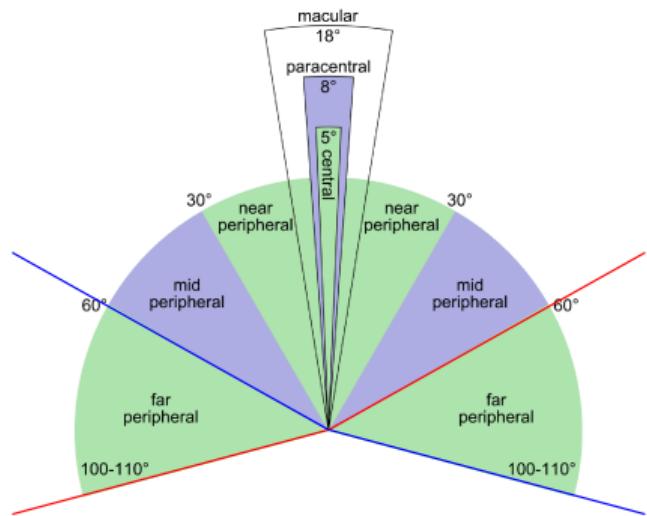
End of Module 02-04

Biological Vision and Applications

Module 02-05: Peripheral Vision

Hiranmay Ghosh

Foveal Vision and Peripheral Vision



- $\approx 99\%$ of visual field is covered by peripheral vision
- Provides an approximate description of the visual field

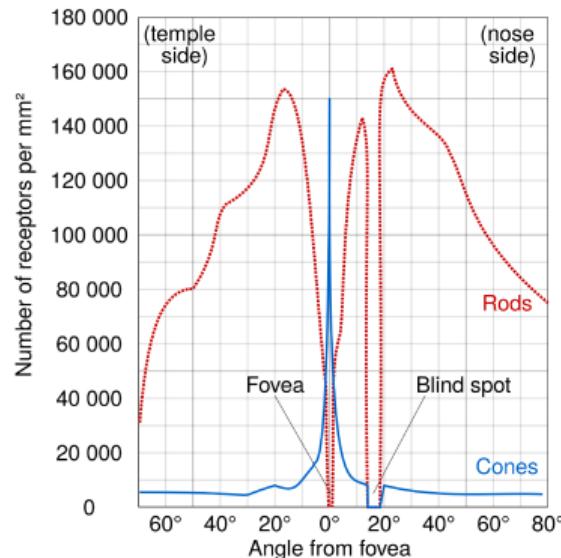
Role of preipheral vision

- Controls eye movement (saccade) in visual search
- Shifts attention to desired place in image



- A more realistic representation of the peripheral view

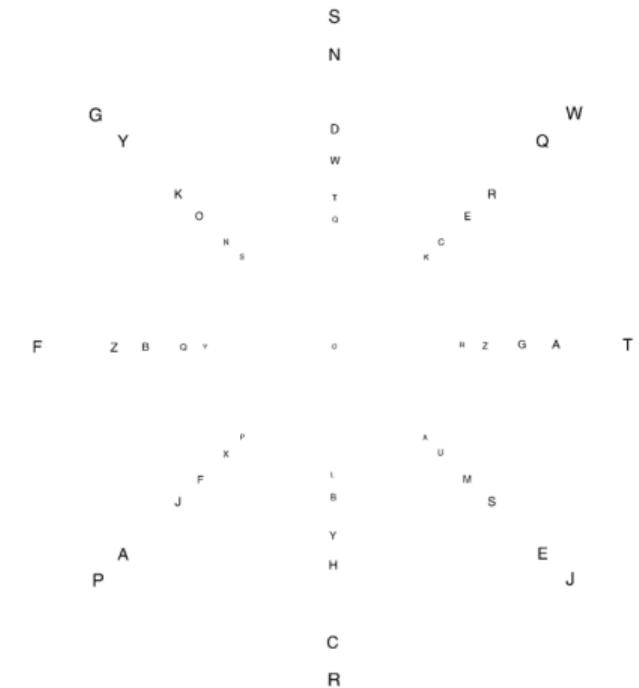
Cortical magnification



- 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

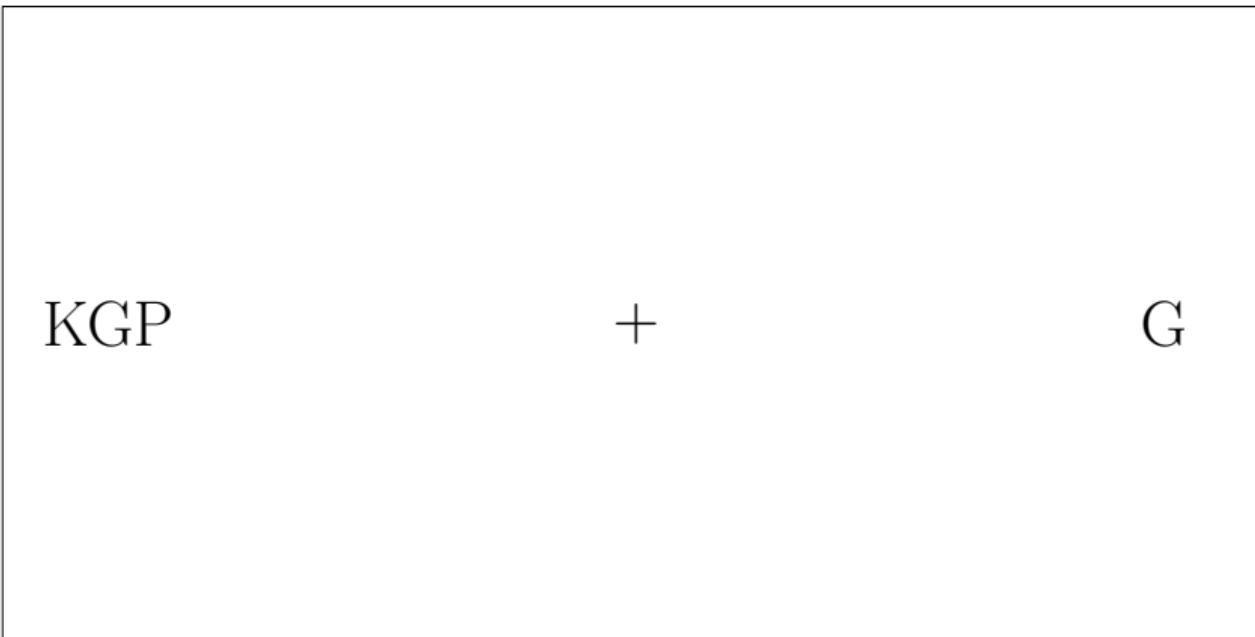
Minimum size of recognizable objects get bigger



Effect of Cortical magnification

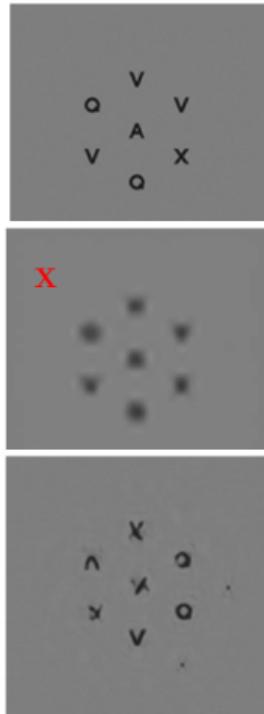
Crowding

- Focus on the cross-hair.
- Try to see the letter 'G' in the left image, and in the right image



Model of Cortical magnification

Some distinctive textures are retained



- Image compression results in blurred image
 - ▶ Pooling / wavelet decomposition
- Cortical magnification does something different
 - ▶ 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
 - ▶ We skip the details

Portilla. A parametric texture model ...

Mongrels



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

Object recognition vs. scene recognition



- Object recognition
- Foveal vision
- Scene recognition
- Peripheral vision

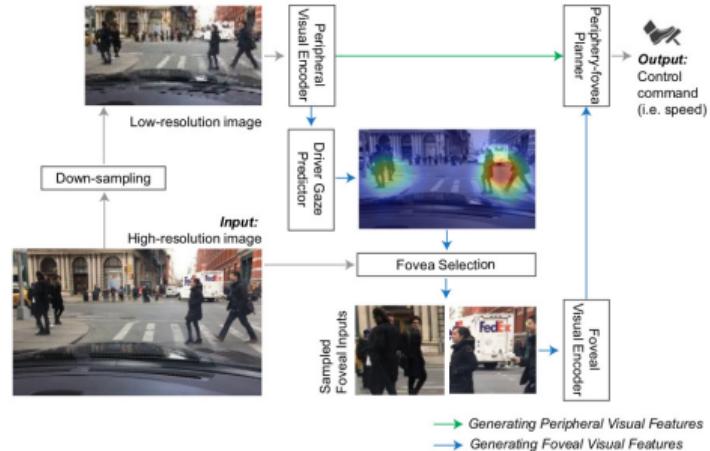
Application: Autonomous Driving



(a) Baseline: all LEDs turn on

(b) Static light: half of the LEDs turn on to hint steering direction

(c) Moving light: a number of LEDs iteratively turn on to hint steering direction



Borojeni, et al. Assisting Drivers with Ambient Take-Over Requests ... (2016)

Xia, et al. Periphery-Fovea Multi-Resolution Driving Model ... (2020)

Application: Logo design

Full-Field View of Logo Designs

Undistorted 512 x 512 Image



Peripheral View of Logo Designs

Foveating the Left-Most Point (x=0, y=256)



Quiz

No quiz for module 02-05

End of Module 02-05

Biological Vision and Applications

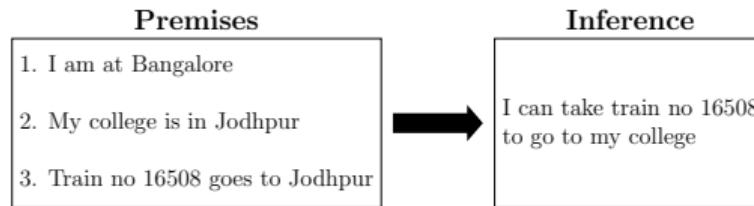
Module 03-01: Reasoning Paradigms

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:



One more example



Premises

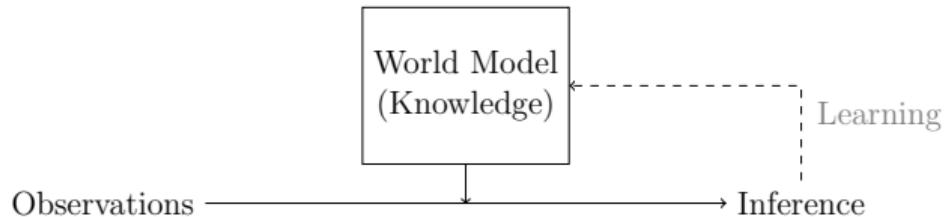


Inference

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

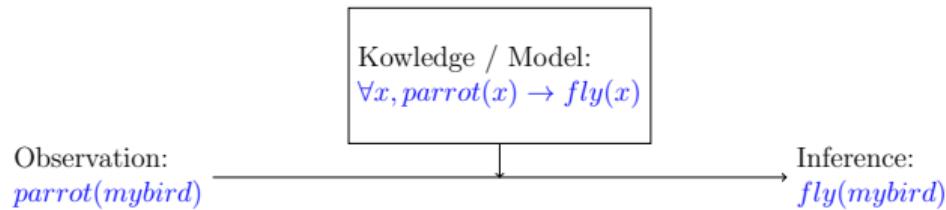


- Create a model of the **world of discourse** (knowledge)
- Interpret observations with that model leading to inference
- **Learning:** Inference may lead to change in the model

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

Rule-based reasoning

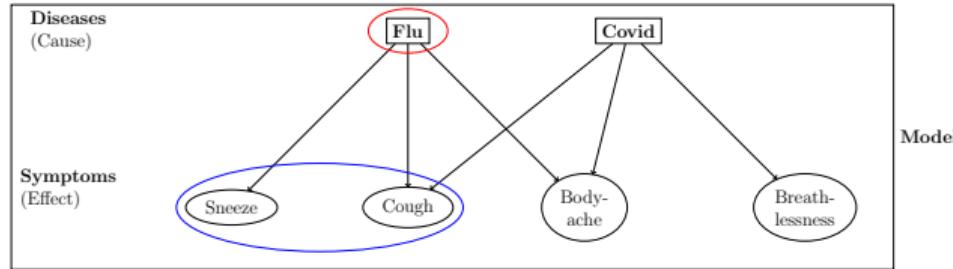
Also called deductive reasoning



- 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 explanation 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

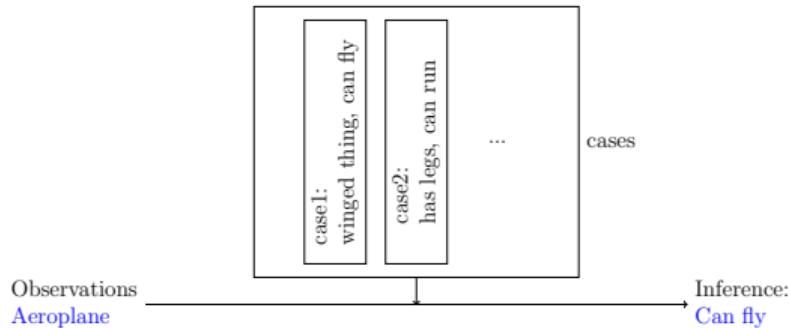
Generalization from observations

- 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 Hierarchical 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

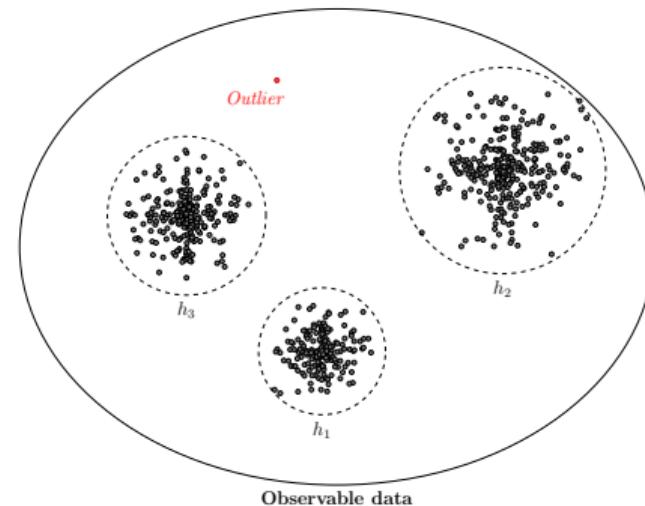
Exact match vs. Inexact match

- 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 → disease 1
- Patients 3,4,5,9 and 10 have similar symptoms → 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 is used in the human mental processes ?
 - ▶ Probably all of them, depending on context
- Which form of reasoning is 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

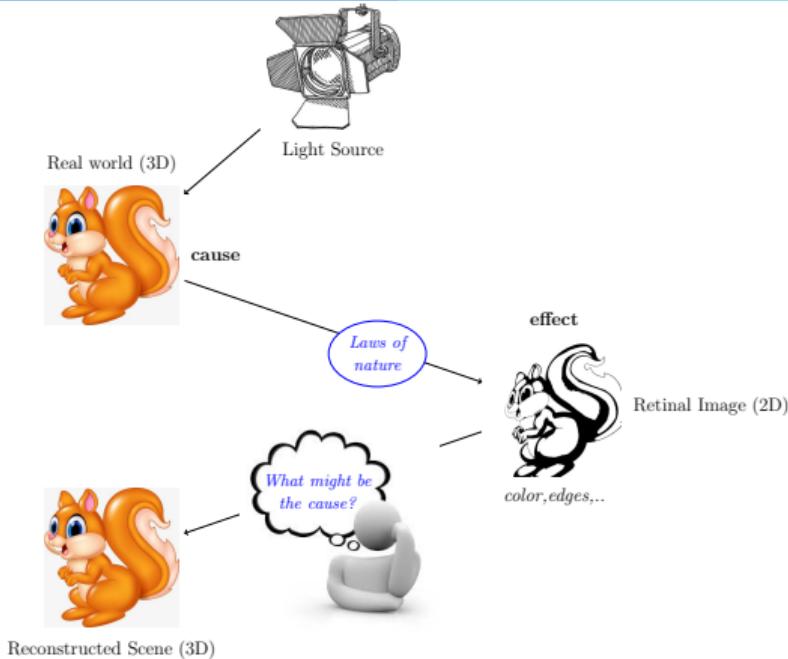
End of Module 03-01

Biological Vision and Applications

Module 03-02: Reasoning for Vision

Hiranmay Ghosh

Vision is an “inverted problem”



- Naturally suited for abduction

Diversity in the natural world

Challenge for Computer Vision

Each human face is different



- How do we recognize a new human face?



Statistical similarity



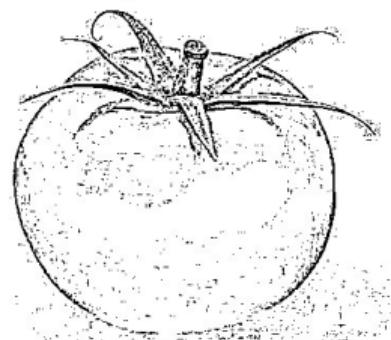
- Fortunately, the images exhibit strong statistical similarity
- Super-imposition of several natural images
 - ▶ scale and pose normalized
 - ▶ Does not result in a blur background
 - ▶ Some statistical features stand out
- Possible to construct a statistical model (object/scene)

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

Imperfection in signal processing

Early vision

- 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 object recognition



Sparsity in image space for natural images

- **Natural scenes:** images captured with devices operating in the range of visual spectrum.
 - ▶ Includes scenes of natural and man-made objects
 - ▶ Excludes text images, computer graphics, animations, paintings, cartoons, X-ray images, etc.
- Combinatorially, it is possible to have $w \times h \times d$ distinct images
 - ▶ w, h : width and height of the image
 - ▶ d : number of possible color values
 - ▶ For a 1024×1024 gray-scale image, we have 256 million possibilities
- All the possibilities never arise in natural images
 - ▶ An image can be represented as a point in an image space with volume $w \times h \times d$
 - ▶ Natural scenes are localized in very narrow regions in the image space

Vision as statistical interpretation



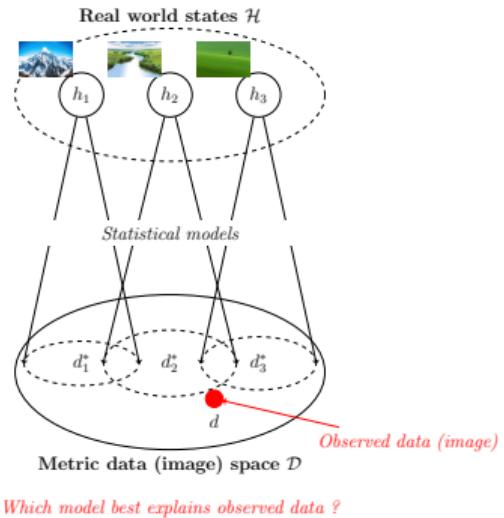
- Human eyes have adapted to the statistics of the natural scenes
 - ▶ Key to robust vision despite noisy image data
 - ▶ e.g., we can “intuitively” reconstruct the occluded contour of the flower
- The statistical regularity is exploited to model vision as a process of statistical interpretation
 - ▶ Robust to natural variations / imperfections

Feature based representation in Computer Vision

- Image space ($w \times h \times d$) is sparse
 - ▶ Scope for compressed representation
- A feature is an abstraction that characterizes the visual contents
 - ▶ It is a lower dimensional representation of an image
 - ▶ Results in data compression
- Examples of features
 - ▶ statistics of edges, colors, etc.

Abductive reasoning for vision

- The real world is in a conceptual state $h_i \in \mathcal{H}$
- A state manifests itself in observable data space \mathcal{D}
 - ▶ Each state has h_i has a statistical model d_i
- We observe some data d
- Which model best explains the observed data?
- Use statistical match – possible because
 - ▶ Statistical similarity of concepts
 - ▶ Sparsity of data space



Quiz



Quiz 03-02

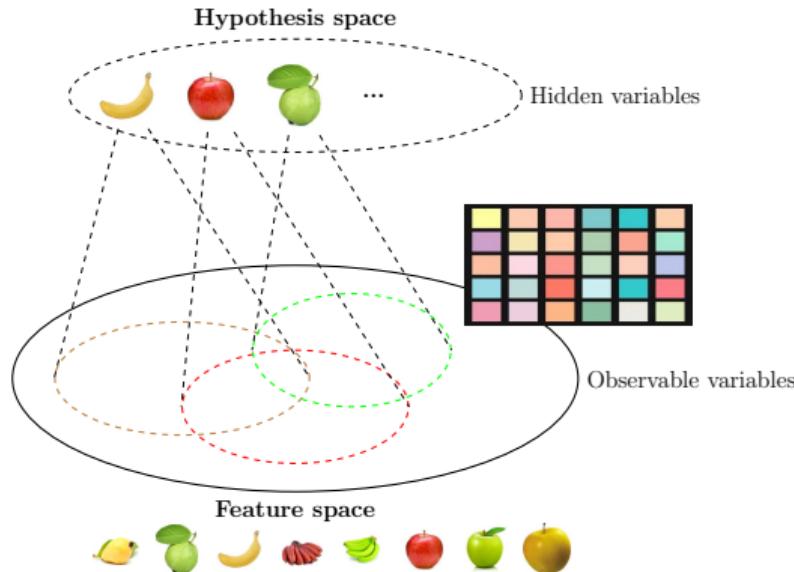
End of Module 03-02

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:

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

$$P(A | B) = \frac{1}{\kappa} \cdot P(B | A) \cdot P(A)$$

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

- 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

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

Using Baye's Theorem:

$$\begin{aligned} P(\text{banana} | \text{yellow}) &= \frac{P(\text{yellow} | \text{banana}) \cdot P(\text{banana})}{P(\text{yellow})} \\ &= \frac{\frac{0.42}{0.7} * 0.7}{0.5} = 0.84 \end{aligned}$$

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			

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

Posterior

Priors

$$\begin{aligned} P(banana \mid yellow) &= \frac{1}{\kappa} * P(yellow \mid banana).P(banana) \\ &= \frac{1}{\kappa} * 0.6 * 0.7 = \frac{1}{\kappa} * 0.42 \end{aligned}$$

How do we get to know the priors and κ ?

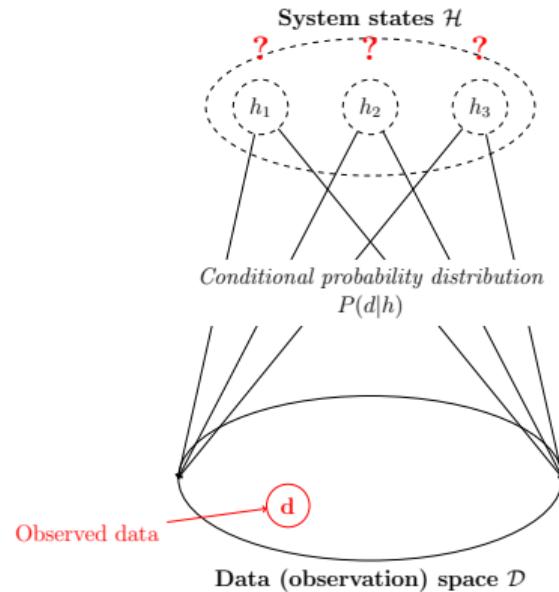
- Priors: $P(banana)$, $P(yellow | banana)$
 - ▶ From external sources / context
 - ▶ From prior observations
- Proportionality constant κ
 - ▶ We do not care
 - ▶ We need to find the **best explanation**
 - ▶ $P(banana | yellow) = \frac{1}{\kappa} * 0.42$
 - ▶ $P(apple | yellow) = \frac{1}{\kappa} * 0.06$
 - ▶ Banana is a better explanation than apple for observed yellow color

Bayesian Inference

Summary

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

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

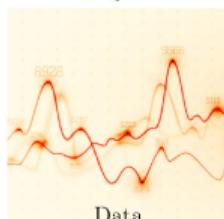


Belief Revision

Prior belief and evidence



Inference



$$\text{Baye's Theorem: } P(h_i | \mathbf{d}) = \frac{1}{\kappa} \cdot P(h_i) \cdot P(\mathbf{d} | h_i)$$

Prior belief

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

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

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

* log base assumed to be 10

- Useful for comparing the plausibility of pairs of concepts

Composite data

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

		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			

$$P(\text{Banana}) = 0.7, P(\text{Apple}) = 0.2, P(\text{Others}) = 0.1$$

- $\mathcal{D} = \{(Red, Long), (Red, Round), \dots\}$
 - ▶ Conditionals: $P(\text{Red, Long} \mid \text{Banana}), \dots$
 - ▶ 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

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

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

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

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

- $P(h_i | \mathbf{d}) = \frac{1}{\kappa} \cdot P(h_i) \cdot P(\mathbf{d} | h_i)$
- Assume that $\mathbf{d} = d_1, d_2, d_3, \dots$ represents a data stream (possibly infinite)
- After d_1 arrives
 - ▶ Posterior: $P(h_i | d_1) = \frac{1}{\kappa_1} \cdot P(h_i) \cdot P(d_1 | h_i)$
 - ▶ This posterior becomes the prior for the second observation
- After d_2 arrives
 - ▶ Posterior: $P(h_i | d_1, d_2) = \frac{1}{\kappa_2} \cdot P(h_i | d_1) \cdot P(d_2 | h_i) = \frac{1}{\kappa_{12}} \cdot P(h_i) \cdot P(d_1 | h_i) \cdot P(d_2 | h_i)$
 - ▶ This posterior becomes the prior for the third observation
- ... and so on
- System updates its 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			

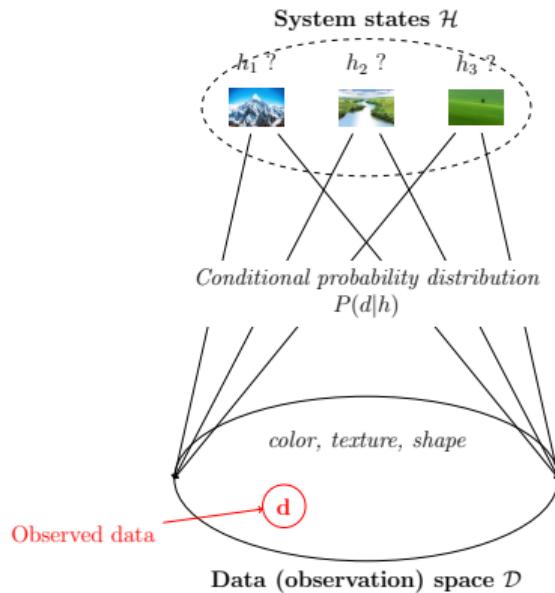
		Fruits (A)				
		Banana	Apple
Shape	Long	0.8	0.3			
	Round	0.2	0.7			
	Total	1	1			

$$P(\text{Banana}) = 0.7, P(\text{Apple}) = 0.2, P(\text{Others}) = 0.1$$

- We see a green and long fruit.
 - ▶ Is it a banana or an apple ?
 - Solution sketch:
 - ▶ Start with prior beliefs $P(\text{banana})$ and $P(\text{apple})$
 - ▶ “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 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
 - ▶ 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: Bayesian Networks

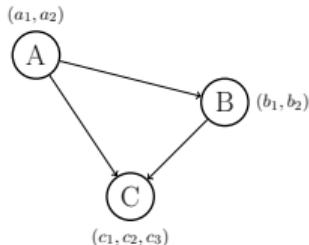
Hiranmay Ghosh

Bayesian reasoning revisited

- Bayesian framework of reasoning
 - ▶ Create a system model in terms of n stochastic (random) variables
 - ▶ $\mathcal{X} = \{X_1, X_2, X_3, \dots, X_n\}$
 - ▶ A variable X_i can have k_i states. $X_i : \{x_i^1, x_i^2, \dots, x_i^{k_i}\}$
 - ▶ 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
- A trivial system with 10 binary variables will have $2^{10} - 1 = 1023$ entries
 - ▶ That is a big number !

Joint probability and conditional probability

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



Non-circular dependency 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.

$$P(a_1, b_1, c_1) = P(a_1, b_1) \cdot P(c_1 | a_1, b_1)$$

$$= P(a_1) \cdot P(b_1 | a_1) \cdot P(c_1 | a_1, b_1)$$

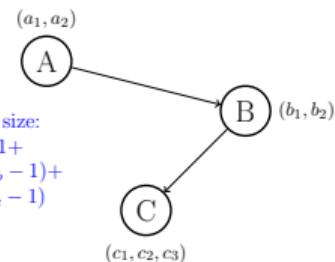
Conditional Independence

Variables A and B are conditionally independent of each other, iff $P(A.B) = P(A).P(B)$

Variables A and B are conditionally independent of each other given C , iff $P(A.B | C) = P(A | C).P(B | C)$

Conditional probabilities

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



This topology assumes A and C are conditionally independent, given B

- Many of the variables in real world are conditionally independent of each other given the state of some other variables ☺, e.g.,
 - ▶ Color of a fruit and its shape, given the fruit
 - ▶ ...
- Conditional independence simplifies probability computations
 - ▶ Another reason to work with conditional probabilities

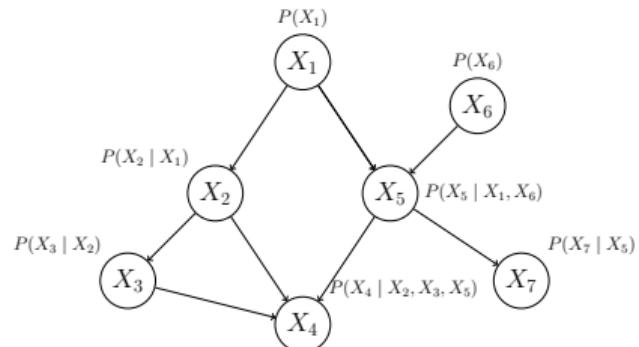
Probabilistic Graphical Models

- 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)

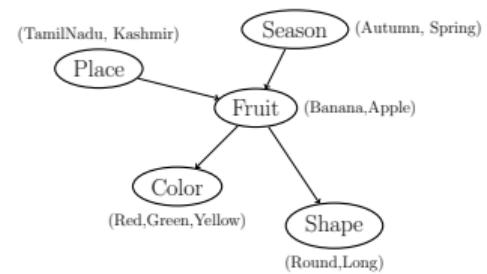
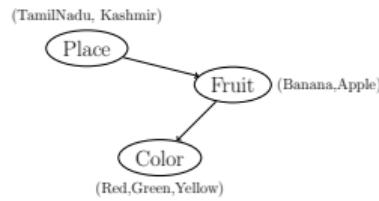
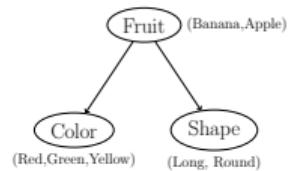
Models a probabilistic reasoning problem with cause-effect relationship



CPT Size:
 $k_1 - 1 +$
 $k_1.(k_2 - 1) +$
 $k_2.(k_3 - 1) +$
 $k_2 k_3 k_5.(k_4 - 1) +$
...

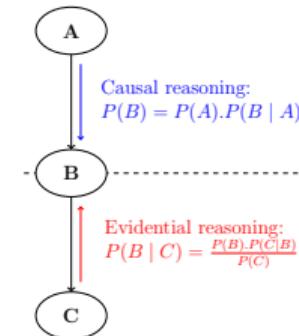
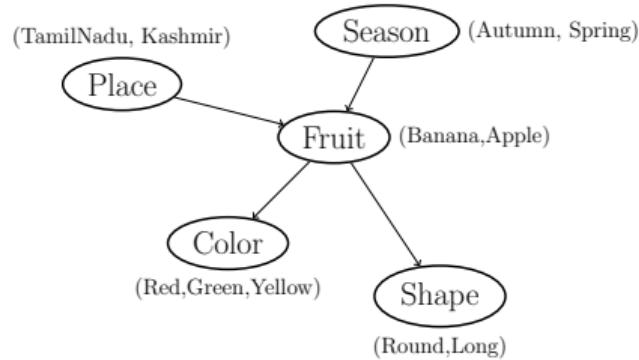
- A Directed Acyclic Graph (DAG)
- Nodes represent events in a system
 - ▶ $X_i = (x_i^1, x_i^2, \dots, x_i^{k_i})$
 - ▶ Some nodes are observable
 - ▶ Others need to be inferred
- Edges represent **causal** relations between the events
- Conditional probabilities $P(X_i | \text{Pa}(X_i))$ are associated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

Examples of BN



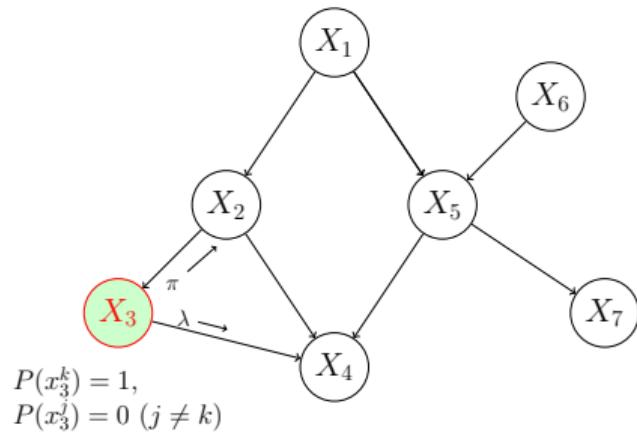
- Naïve Bayesian Network

Causal inference and Evidential inference



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

Inferencing with Bayesian Networks

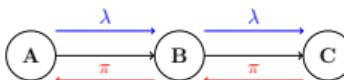


- Hand compute probabilities
 - ▶ There can be multiple (undirected) paths between a pair of nodes
 - ▶ Extremely complex
- Pearl's belief propagation algorithm
 - ▶ π and λ messages
 - ▶ Probabilities of neighboring nodes updated
 - ▶ Traverses recursively in the network
 - ▶ Till no more nodes left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

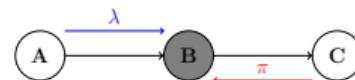
Pearl's algorithm

Network topology and Belief propagation

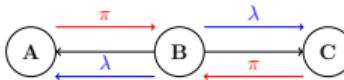
D-Separation



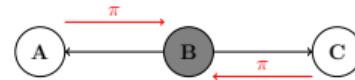
A causes B, B causes C, State of B is unknown
Belief flows between A and C



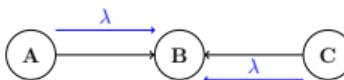
A causes B, B causes C, State of B is known
The path between A and C is blocked by B



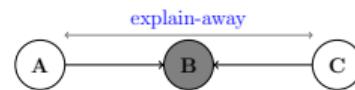
B is the cause of A and C, state of B is unknown
Belief flows between A and C



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

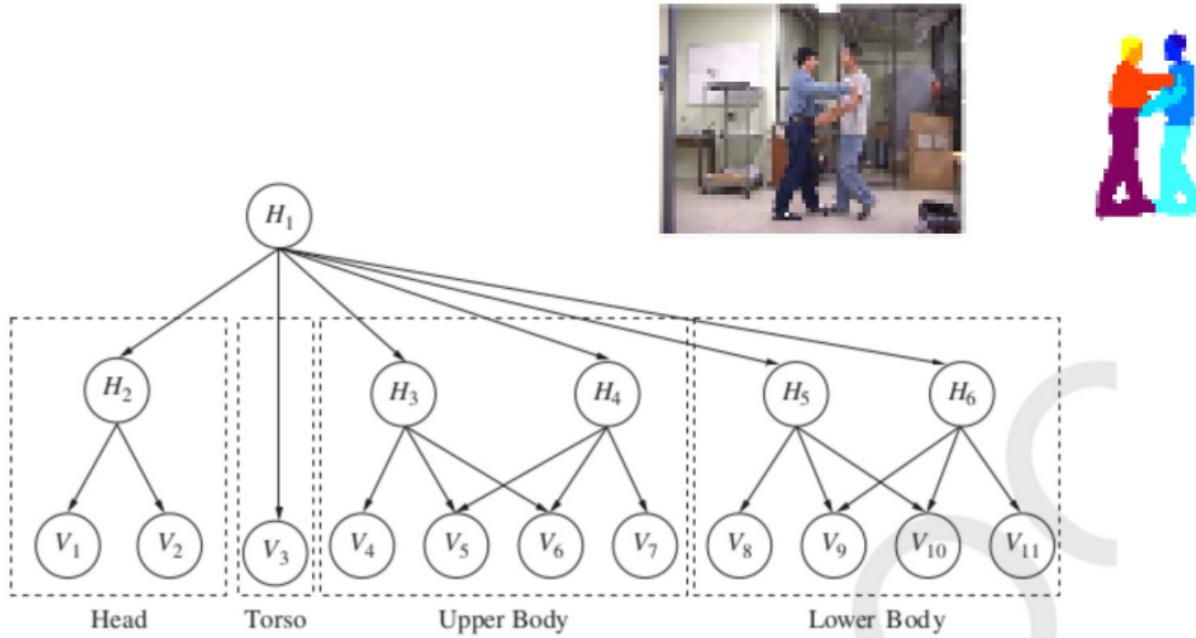


A and C are causes of B, state of B is known
A explains away C, and vice-versa

- 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 d-separated

Hierarchical organization in Bayesian Network

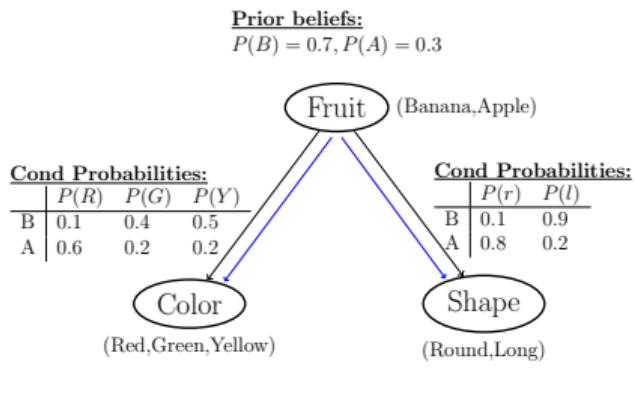
Example



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

Simple Bayesian Network Example

Prior beliefs, Conditional probabilities and likelihoods



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

Colors:

$$\begin{aligned}P(\text{Red}) &= P(R | B) \times P(B) + P(R | A) \times P(A) = 0.25 \\P(\text{Green}) &= 0.34 \\P(\text{Yellow}) &= 0.41\end{aligned}$$

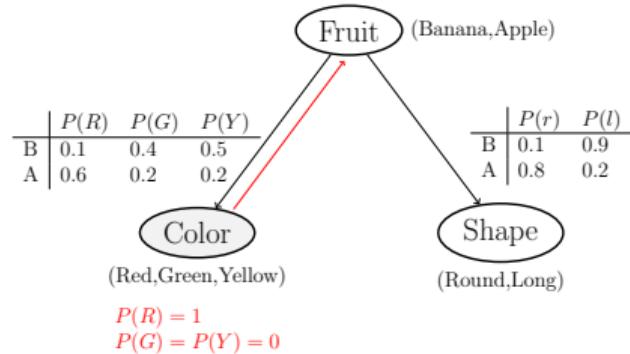
Shapes:

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

Simple Bayesian Network Example

Posteriors

$$P(B) = 0.7, P(A) = 0.3$$



$$\begin{aligned}P(R) &= 1 \\P(G) &= P(Y) = 0\end{aligned}$$

- We see a fruit to be red

Fruits: (un-normalized)

$$\begin{aligned}P(\text{Banana} \mid \text{Red}) &= P(R \mid B) \times P(B) \\&= 0.1 \times 0.7 = 0.07 \\P(\text{Apple} \mid \text{Red}) &= 0.18\end{aligned}$$

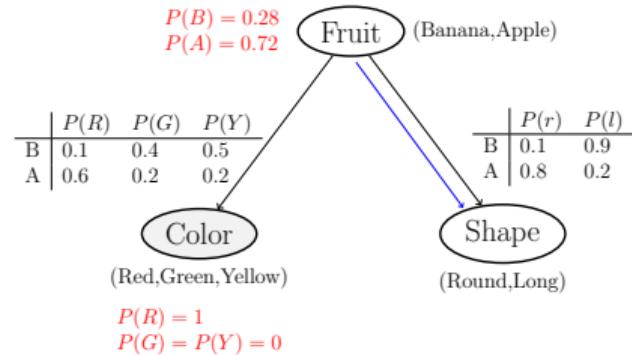
Fruits: (normalized)

$$\begin{aligned}P(\text{Banana} \mid \text{Red}) &= 0.28 \\P(\text{Apple} \mid \text{Red}) &= 0.72\end{aligned}$$

Simple Bayesian Network Example

Posteriors (contd.)

$$P(B) = 0.7, P(A) = 0.3$$



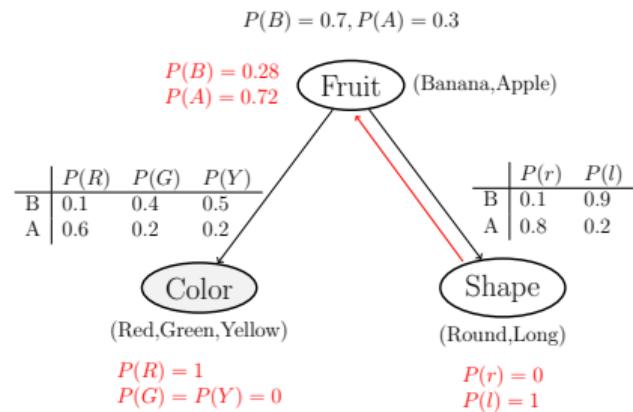
- Change in probability of fruits changes posterior probability of shapes

Shapes:

$$\begin{aligned}P(\text{Round}) &= 0.60 \\P(\text{Long}) &= 0.40\end{aligned}$$

Simple Bayesian Network Example

Posteriors (contd.)



- Now we see the fruit to be **long**

Fruits: (un-normalized)
 $P(\text{Banana} \mid \text{Red}, \text{Long}) = P(l \mid B) \times P(B)$
 $= 0.9 \times 0.28 = 0.25$
 $P(\text{Apple} \mid \text{Red}, \text{Long}) = 0.14$

Fruits: (normalized)
 $P(\text{Banana} \mid \text{Red}, \text{Long}) = 0.63$
 $P(\text{Apple} \mid \text{Red}, \text{Long}) = 0.37$

Quiz



Quiz 03-04

End of Module 03-04

Biological Vision and Applications

Module 03-05: 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?
 - ▶ Domain theory
 - ▶ Past observations
- From your past observations
 - ▶ You observed 20 bananas; 2 of them are red
 - ▶ $P(\text{red} \mid \text{banana}) = \frac{2}{20} = 0.1$

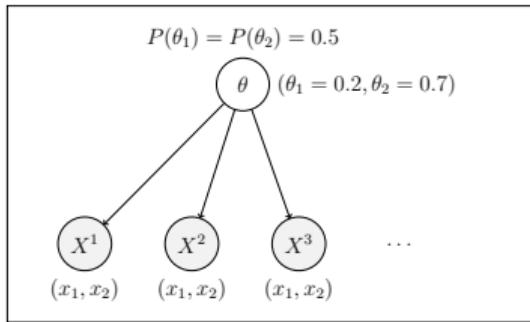
- Maximum Likelihood Estimation (MLE)
 - ▶ Let X be a stochastic variable with n possible values: x_1, \dots, x_n
 - ▶ We make N experiments
 - ▶ Observe r_i occurrences for $X = x_i$ $(\sum_{i=1}^n r_i = N)$
 - ▶ Estimates of $P(X = x_i) = \frac{r_i}{N}$

Maximum Likelihood estimation

Pitfalls

- Data-driven approach
- Sparsity of data
- Extremely unreliable, if the sample size is small
 - ▶ Observe 2 bananas, one of them is red: $P(\text{red} \mid \text{banana}) = \frac{1}{2}$
 - ▶ Intuitively, this is incorrect!
- Does not tell how reliable the estimate is
 - ▶ Does not distinguish between (2 out of 20) and (200 out of 2000)
 - ▶ Cannot tell $P(X = x_1) = 0.1 \pm 5\%$

Bayesian Estimation



- Assume that X can have two values: (x_1, x_2)
- Hidden parameter $\theta \equiv P(X = x_1)$:
 - ▷ Causes outcomes of the experiments
 - ▷ $P(X = x_2) = 1 - \theta$
- Given θ , experiments are conditionally independent
- Assume that θ can have two possible values: $(\theta_1 = 0.2, \theta_2 = 0.7)$
- Assume that they are equiprobable to begin with:
 - ▷ $P(\theta_1) = P(\theta_2) = 0.5$ [$P(\theta_1) \equiv P(\theta = \theta_1), P(\theta_2) \equiv P(\theta = \theta_2)$]

$$\begin{aligned}P(X^1 = x_1) &= P(\theta_1).P(x_1 | \theta_1) + P(\theta_2).P(x_1 | \theta_2) \\&= P(\theta_1).\theta_1 + P(\theta_2).\theta_2 = 0.5 \times 0.7 + 0.5 \times 0.2 = 0.45\end{aligned}$$

Now, we make the first experiment. Assume $X^1 = x_1$

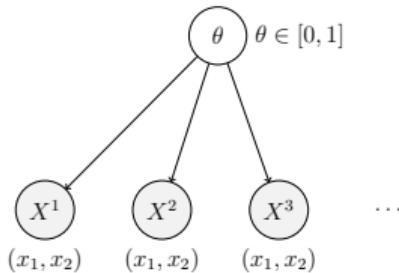
$$P(\theta_1 | X^1 = x_1) = \frac{P(\theta_1).P(X^1=x_1|\theta_1)}{P(X^1=x_1)} = \frac{0.5 \times 0.2}{0.45} \approx 0.22$$

$$P(\theta_2 | X^1 = x_1) = \frac{P(\theta_2).P(X^1=x_1|\theta_2)}{P(X^1=x_1)} = \frac{0.5 \times 0.7}{0.45} \approx 0.78$$

Bayesian Estimation

contd.

Uniform distribution: $p_0(\theta) = 1$



Predicted outcome of 1st experiment:

$$\hat{\theta}_0 = P(X^1 = x_1 | p_0(\theta)) = \int_0^1 p_0(\theta) \cdot \theta \cdot d\theta = \int_0^1 \theta \cdot d\theta = \frac{1}{2}$$

We perform the first experiment. We observe $X_1 = x_1$

$$p_1(\theta) = p(\theta | p_0(\theta), X^1 = x_1) = \frac{p_0(\theta) \cdot P(X^1=x_1|\theta)}{P(X^1=x_1)} = \frac{1 \times \theta}{1/2} = 2\theta$$

Predicted outcome of 2nd experiment:

$$\hat{\theta}_1 = P(X^2 = x_1 | p_1(\theta)) = \int_0^1 p_1(\theta) \cdot \theta \cdot d\theta = \int_0^1 2\theta^2 \cdot d\theta = \frac{2}{3}$$

Now we make the 2nd experiment. We observe $X^2 = x_2$

$$p_2(\theta) = p(\theta | p_1(\theta), X^2 = x_2) = \frac{p_1(\theta) \cdot P(X^2=x_2|\theta)}{P(X^2=x_2)} = \frac{2\theta \times (1-\theta)}{1-2/3} = 6\theta(1-\theta)$$

Predicted outcome of 3rd experiment:

$$\hat{\theta}_2 = P(X^3 = x_1 | p_2(\theta)) = \int_0^1 p_2(\theta) \cdot \theta \cdot d\theta = \int_0^1 2/3 \cdot \theta^2 (1-\theta) \cdot d\theta = \frac{1}{2}$$

We could also compute the result in a single step

▷ sequence of observation does not matter

$$p_2(\theta) = p(\theta | p_0(\theta), X^1 = x_1, X^2 = x_2) = \frac{p_0(\theta) \cdot P(X^1=x_1, X^2=x_2|\theta)}{P(X^1=x_1, X^2=x_2)} = \frac{\theta \times (1-\theta)}{\int_0^1 \theta \cdot (1-\theta) \cdot d\theta} = 6\theta(1-\theta)$$

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

Bayesian Estimation

- Assume that we have made n experiments
 - ▷ $D = \langle x_1, x_2, x_1, x_1, \dots \rangle$
 - ▷ We have observed k x_1 s and $(n - k)$ x_2 s (in whatever sequence)

$$P(D \mid \theta) = \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

$$p_n(\theta) = p(\theta \mid p_0(\theta), D) = \frac{p_0(\theta) \cdot P(D|\theta)}{P(D)} = \frac{\frac{1}{\int_0^1} \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}}{\int_0^1 \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k} \cdot d\theta} = \frac{(n+1)!}{k! \cdot (n-k)!} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

$$\hat{\theta}_n = P(X_{n+1} \mid p_n(\theta)) = \int_0^1 \theta \cdot p_n(\theta) \cdot d\theta = \frac{(n+1)!}{k! \cdot (n-k)!} \cdot \int_0^1 \theta^{k+1} \cdot (1 - \theta)^{n-k} \cdot d\theta$$

$$= \frac{(n+1)!}{k! \cdot (n-k)!} \times \frac{(k+1)! \cdot (n-k)!}{(n+2)!} = \frac{k+1}{n+2}$$

We have used the result $\int_0^1 x^m \cdot (1 - x)^n \cdot dx = \frac{m! \cdot n!}{(m+n+1)!}$

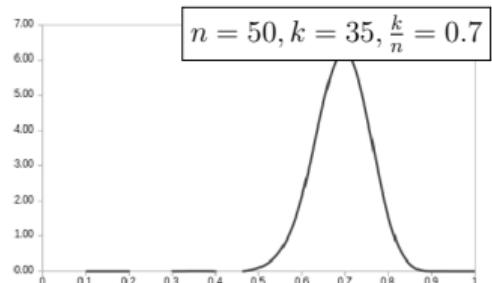
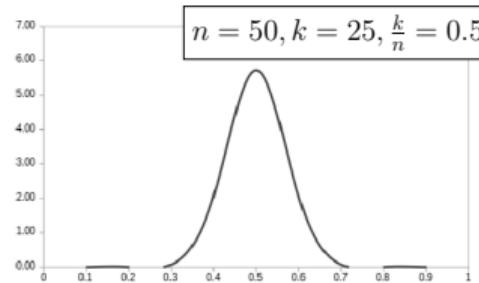
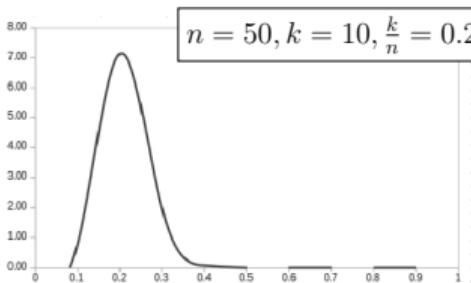
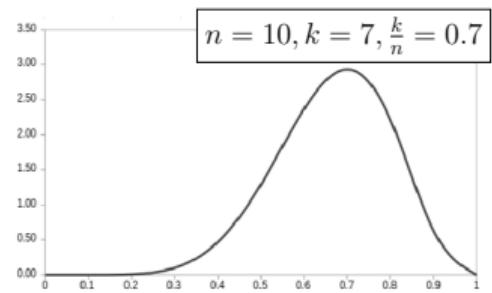
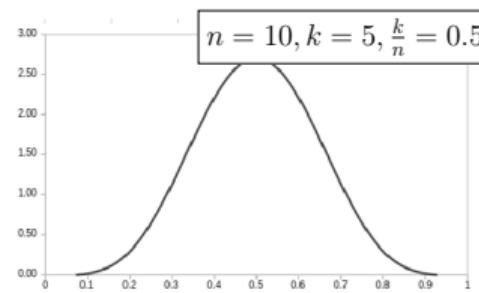
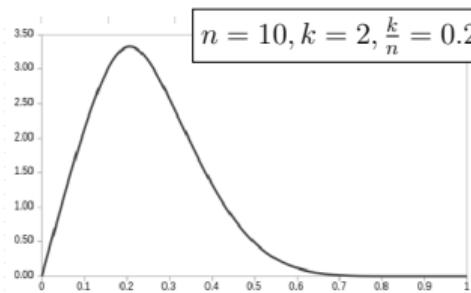
Discussions

- $p_0(\theta) = 1$ [Uniform distribution]
- $D = \langle x_1, x_2, x_1, x_1, \dots \rangle$ [k x_1 s and $(n - k)$ x_2 s in any sequence]
- $p_n(\theta) = p(\theta | p_0(\theta), D) = \frac{(n+1)!}{k!(n-k)!} \cdot \theta^k \cdot (1 - \theta)^{n-k}$
- $\hat{\theta}_n = \frac{k+1}{n+2}$

- We get a pdf for θ , rather than a single value
 - ▶ More informative, spread tells how reliable it is
- Expected value for θ is similar to MLE
 - ▶ ... with two additional experiments with outcomes x_1 and x_2 respectively
- Prior belief in Bayesian estimate moderates the extreme observations
 - ▶ Let's assume, we have observed 2 bananas, none is red ($n = 2, k = 0$)
 - ▶ MLE: $\theta = \frac{k}{n} = 0$, Bayesian: $\hat{\theta} = \frac{k+1}{n+2} = \frac{1}{4}$
- Bayesian estimation approaches MLE for large n

Dependence of pdf on data

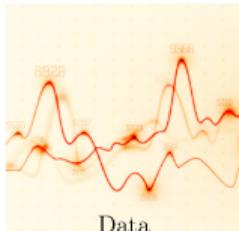
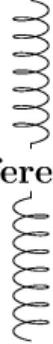
Assumes uniform prior belief [$p_0(\theta) = 1$]



Priors vs. data (observations)



Inference



- We have assumed uniform pdf $p(\theta) = 1$ in this example.
- It is possible of assume other priors
- What determines the priors?
 - ▶ Theory or explanations
 - ▶ Observation in other domains and inductive generalization
- Cognitive bias

Quiz



Quiz 03-05

End of Module 03-05

Biological Vision and Applications

Module 03-06: Occam's razor

Hiranmay Ghosh

Occam's Razor

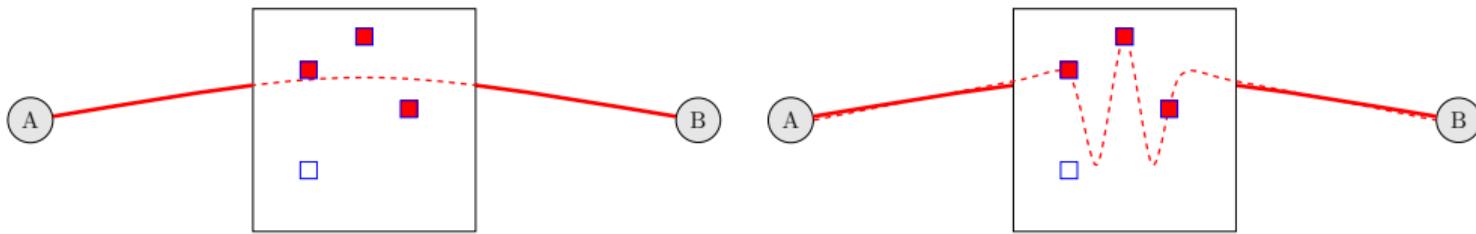
Human mind tends to choose the simplest explanation



- Intuitively, which of the possibilities will you choose ?

Occam's razor

... the observations should be explained

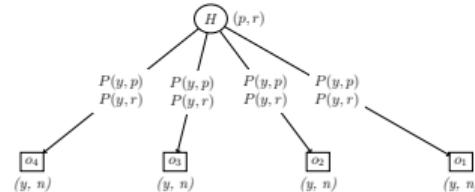
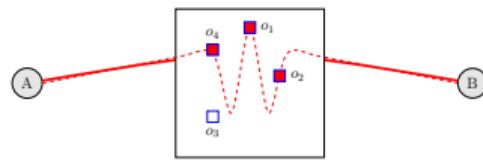


- Which of the possibilities do you choose now ?
- Inference is a tradeoff between complexity of model and **goodness of fit**
 - ▶ Goodness of fit: How well does the model explain the data

EdPuzzle – Occam's razor

Complexity of hypothesis & Goodness of fit

Complexity of hypothesis: $C(p) = \text{No. of parameters required to define the curve}$



Goodness of fit: $GoF = \frac{P(p|D)}{P(r|D)}$

Complexity of explanation: $C(p | D) = -\log_2(GoF)$

Complexity and belief

- Let $c(M)$ denote the complexity for a proposition M
- Prior belief in model M monotonically decreases with complexity
 - We assume an exponential model: $P(M) = 2^{-c(M)}$, $c(M) = -\log_2 P(M)$

- $c(h_i) = -\log_2 P(h_i)$: complexity of the hypothesis (prior)
- $c(d | h_i) = -\log_2 P(d | h_i)$: complexity of evidence given the hypothesis
(inverse of goodness of fit)
- $c(h_i | d) = -\log_2 P(h_i | d)$: complexity of the inference

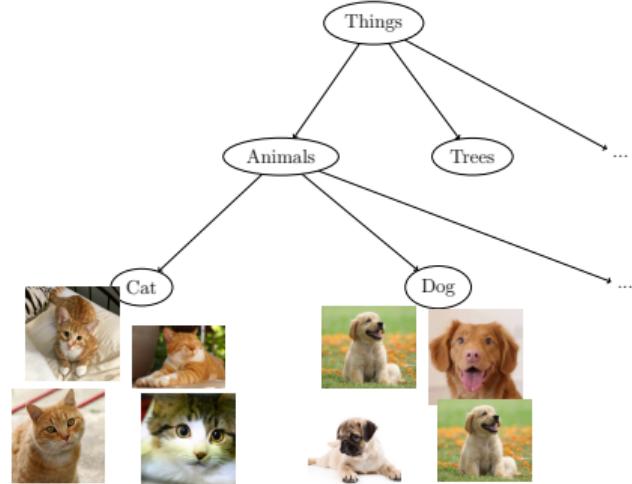
- Substituting in Baye's theorem $P(h_i | d) = \kappa \cdot P(h_i) \cdot P(d | h_i)$
 - $c(h_i | d) = k + c(h_i) + c(d | h_i)$

Belief maximization \equiv complexity minimization

- Bayesian inference: $h^* = \operatorname{argmax}_i P(h_i | d) = \operatorname{argmax}_i P(h_i).P(d | h_i)$
- Equivalently: $h^* = \operatorname{argmin}_i c(h_i | d) = \operatorname{argmin}_i [c(h_i) + c(d | h_i)]$
- Human mind chooses the inference with least complexity
- Inference is a tradeoff between simplicity of the prior and goodness of fit

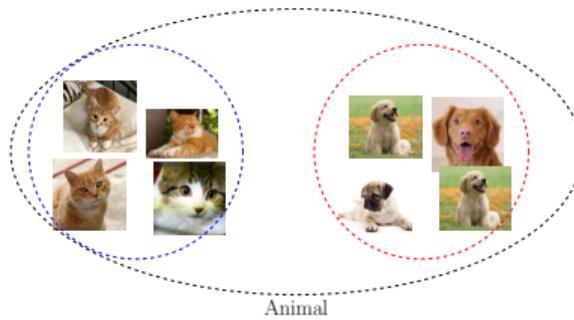
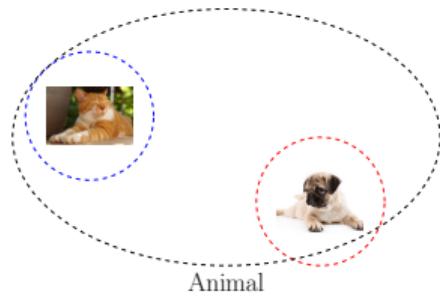
Taxonomy

Organizing concepts in a hierarchy



- Learned top-down, or bottom-up ?

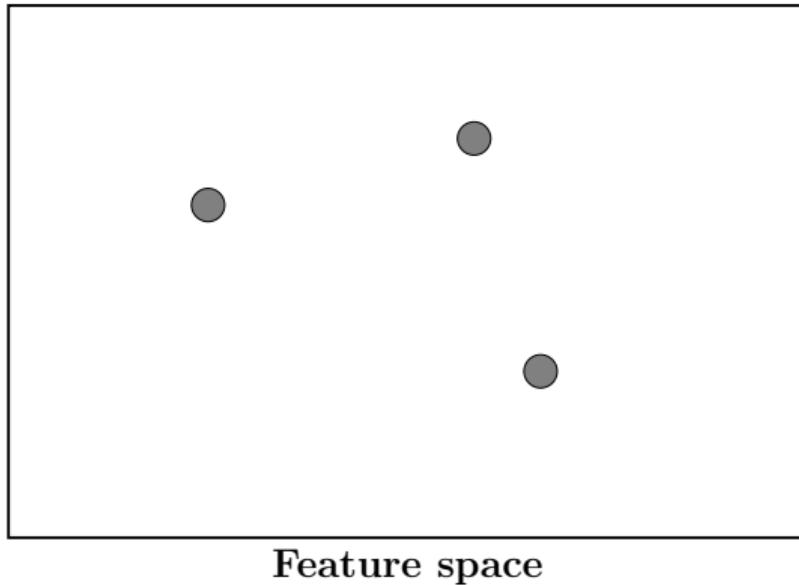
Taxonomy Learning



- Taxonomy is a tradeoff between complexity of hypothesis (number of classes) and goodness of fit

Example

Sparse data



Hypothesis 1: One class

Simple hypothesis

Poor goodness of fit

Hypothesis 2: Three classes

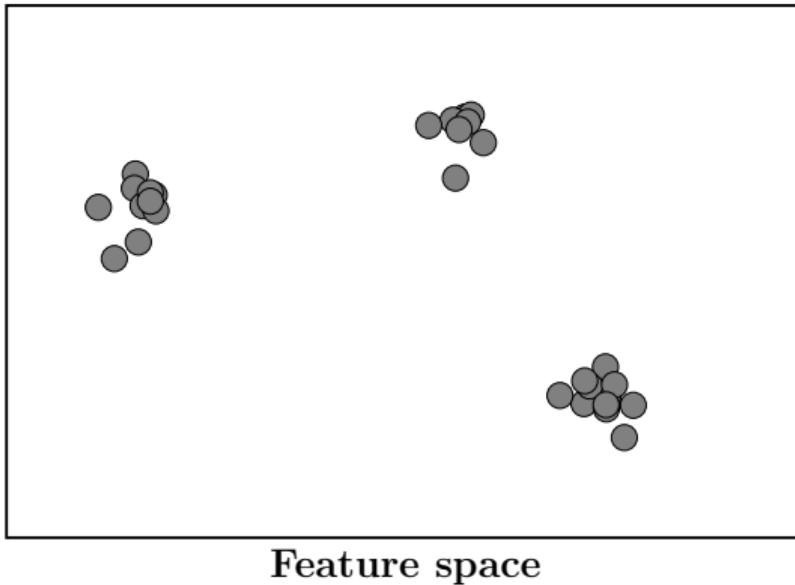
Complex hypothesis

Better goodness of fit

- Intuitively, which one is more acceptable ?

Example

Dense data



Hypothesis 1: One class

Simple hypothesis

Poor goodness of fit

Hypothesis 2: Three classes

Complex hypothesis

Better goodness of fit

- Intuitively, which one is more acceptable ?

Bayesian approach to taxonomy learning

- To minimize: $c(h_i | d) = c(h_i) + c(d | h_i)$
- A hypothesis h_i is characterized by
 - ▶ N_i : Number of classes proposed in the hypothesis
 - ▶ where each class is characterized by (n_i, A_i)
 - ▶ n_i : number of data points
 - ▶ A_i : area (tightest fit to all data points)
- Complexity of hypothesis (prior): $c(h_i) = k_1 \cdot N_i$

Bayesian approach to taxonomy learning

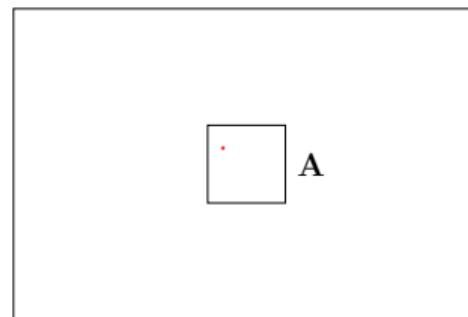
... cont'd.

Probability of a random point to fall in $\mathbf{A} = \frac{A}{F}$
Probability of n random points to fall in $\mathbf{A} = \left(\frac{A}{F}\right)^n$

Goodness of fit: $GF \propto 1/(A^n)$

Data complexity: $\log \frac{1}{GF} = k + n \cdot \log A$

Feature Space \mathbf{F}



- There are N_i classes in hypothesis h_i , each characterized by (n_i, A_i)
 - ▶ Complexity of evidence: $c(d | h_i) = k_2 + \sum_{i=1}^{N_i} n_i \cdot \log A_i$

Bayesian approach to taxonomy learning

... cont'd.

- Complexity of the posterior belief for hypothesis h_i :
 - ▶ Complexity of prior + Complexity of evidence
 - ▶ $c(h_i | d) = k_1.N + k_2 + \sum_{i=1}^N n_i.logA_i$

- Inference: $h^* = \operatorname{argmin}_i c(h_i | d) = \operatorname{argmin}_i (k_1.N + k_2 + \sum_{i=1}^N n_i.logA_i)$

Let us work out one numerical example for better understanding (quiz)

You can refer to this slide while attending the quiz

Quiz



Quiz 03-06

End of Module 03-06

Biological Vision and Applications

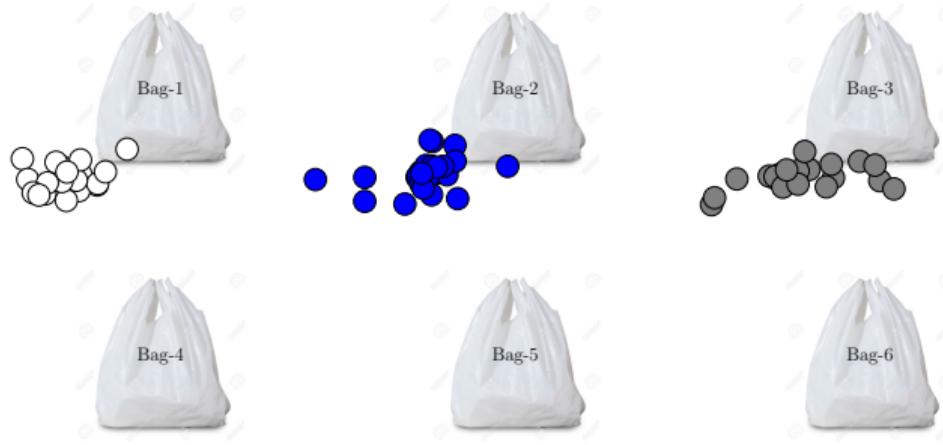
Module 03-07: Hierarchical Bayesian Model



Hiranmay Ghosh

An example

Prior belief: The bags can have marbles of any color or mix



- What do we learn from these observations?
- Can we predict something about bags 4 – 6 that are yet to be sampled?

An example

... Contd.

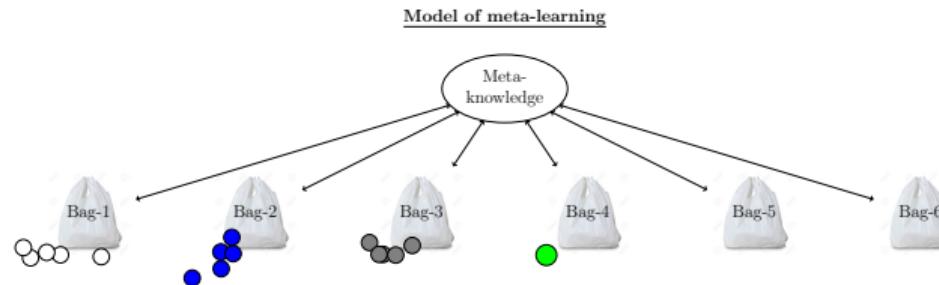


- What do we infer about bag 4 from this new observation?

This is the very basis of transfer learning

Specific knowledge and Generic knowledge

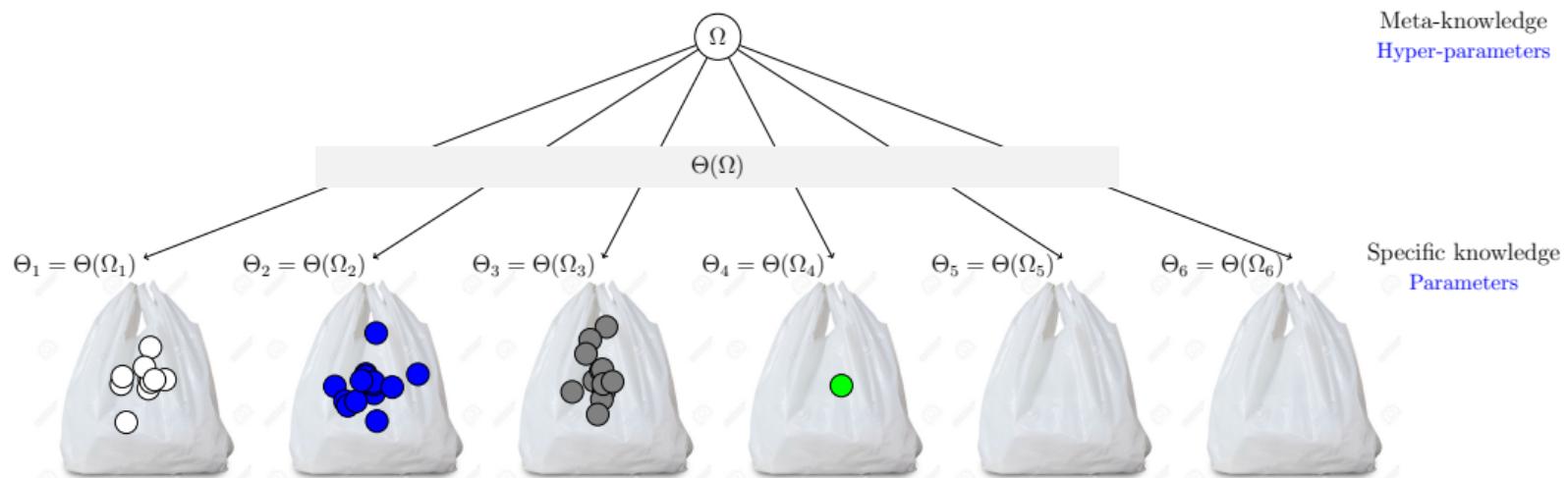
- **Specific Knowledge:** When we sample marbles from a particular bag, we gain knowledge about the content of that bag
- **Generic (Meta) Knowledge:** When we sample marbles from several bags, we gain knowledge about all bags ... even for those which are not sampled



This is an instance of **inductive reasoning** or **inductive generalization**

Modeling the problem

Hierarchical Bayesian Model



Modeling the problem

contd ...

- Let Θ_i represent the model parameters for bag i
 - ▶ $\Theta_i = (\theta_{i1}, \theta_{i2}, \dots)$, θ_{ij} : probability of a marble to be of color j
 - ▶ $0 \leq \theta_{ij} \leq 1, \sum_j \theta_{ij} = 1$
 - ▶ Parameters θ_{ij} 's can be individually learned using Bayesian inferencing
- Θ_i 's are modeled as probabilistic functions of some hyper-parameters Ω in HBM
- A common approach is to use Dirichlet distribution

$$\text{Dirichlet distribution: } P_\alpha(x) = \frac{1}{B(\alpha)} \prod_{i=1}^k x_i^{(\alpha_i+1)}, \quad \text{where } B(\alpha) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)}$$

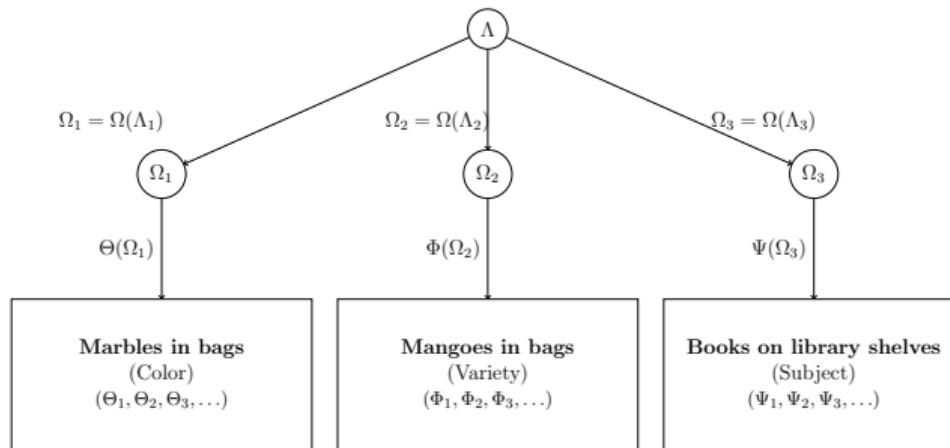
Example of Dirichlet Distribution

What is really happening?

- When we have no observations, we have some priors on Ω
 - ▶ This determines priors (constraints) for Θ_i 's
- As we observe i -th bag, we learn (update) Θ_i
 - ▶ As we “observe” Θ_i , we learn (update) Ω
- As we update Ω , values of all Θ_i are updated

- Hyper-parameters Ω are learned together with the model parameters Θ_i 's
- Hyper-parameters Ω links the model parameters Θ_i 's
- An observation for one bag serves as an observation for the other bags too

Progressive generalization of knowledge



- It is possible to model generic knowledge with even higher abstraction (level) of knowledge, and so on ...
- The entire knowledge-base gets linked
 - ▶ Generalization from one problem to another will be efficient for similar problems

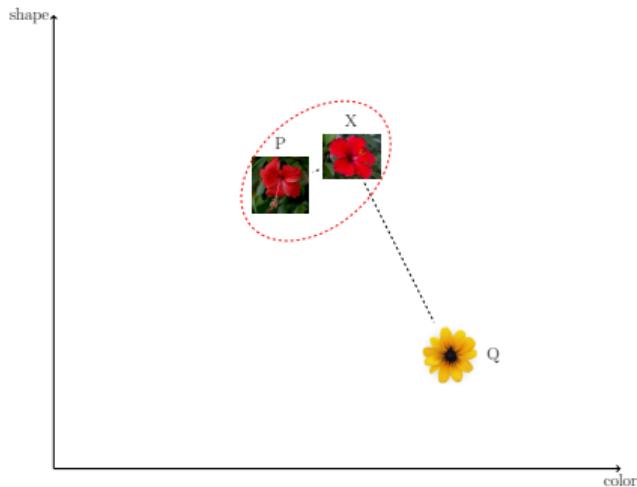
Feature Learning

Which feature do you choose ?



Which category X belongs to?

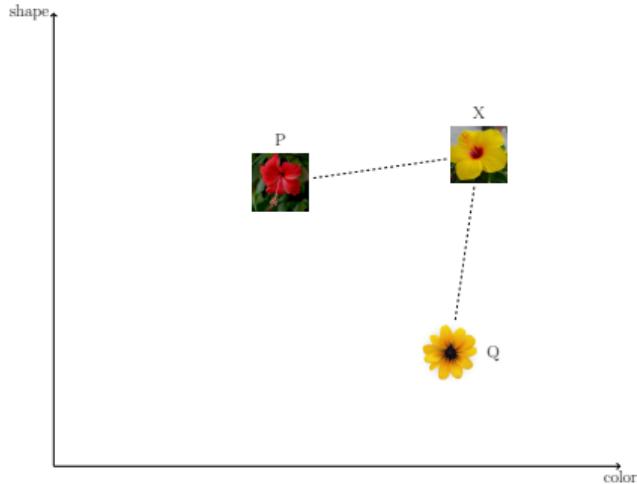
P and Q are rare flowers, you have one sample for each



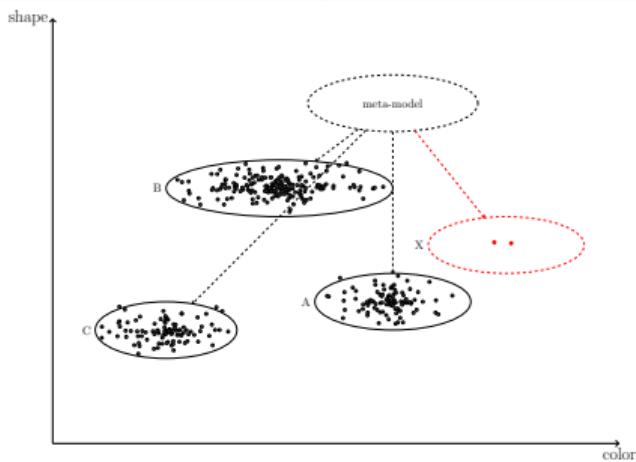
Pretty simple !

Which category X belongs to?

P and Q are rare flowers, you have one sample for each



Meta-learning from abundant classes

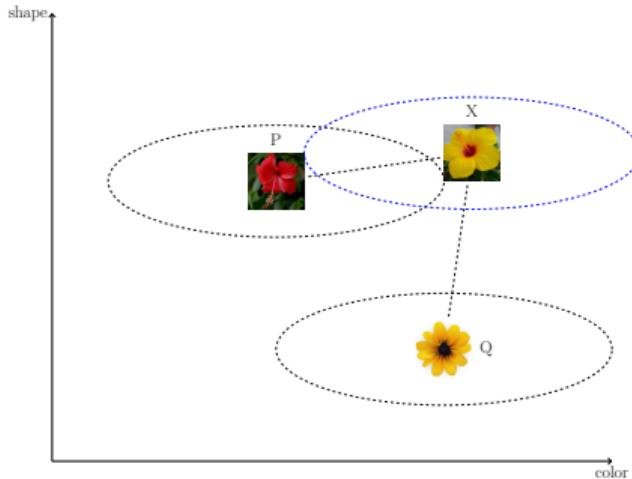


Meta-learn visual model for objects from abundant classes:

Object features have more spread in color than shape

Use the meta-model to create models for new (and rare) classes

Create models for rare classes from meta-model



X belongs to class P

We put more emphasis on shape than color

Shape bias

That is exactly how a child learns to distinguish objects by their shapes



Source: Shape Bias

Quiz



Quiz 03-07

End of Module 03-07

Biological Vision and Applications

Module 04-01: Feature Integration Theory

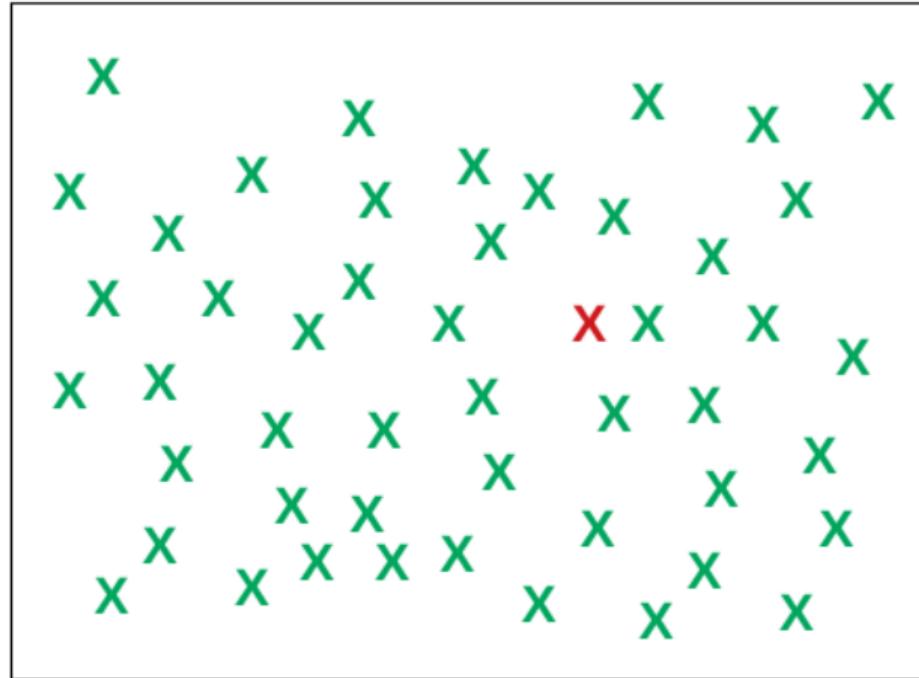
Hiranmay Ghosh

Experiment in visual Search

- We shall show you two slides with one figure each
- There is exactly one **red X** in each of the figures, besides other characters
- You will have to find the **red X** in the figures

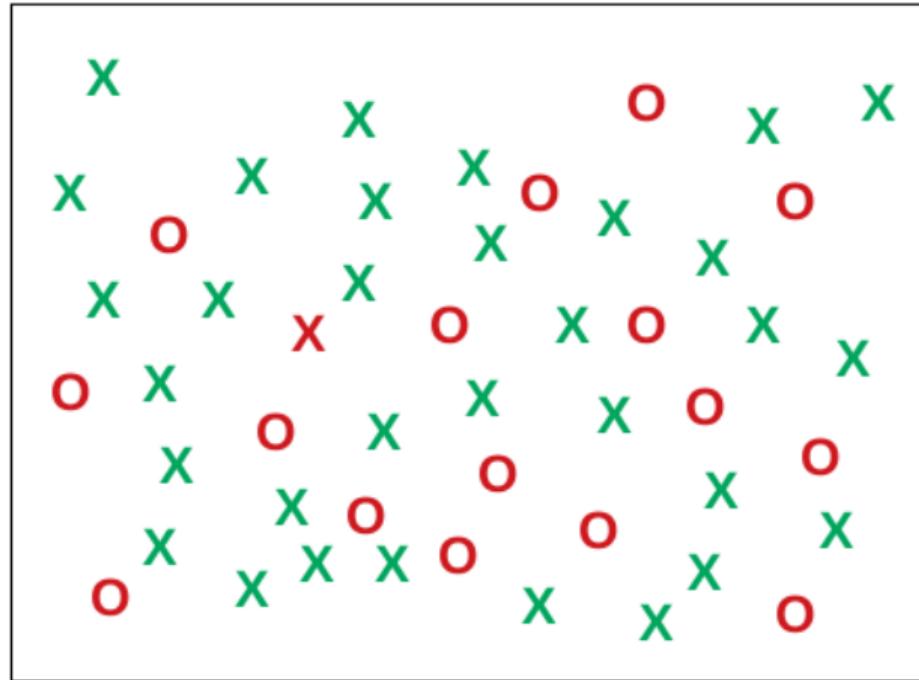
Experiment in visual Search

Find the red X in the figure



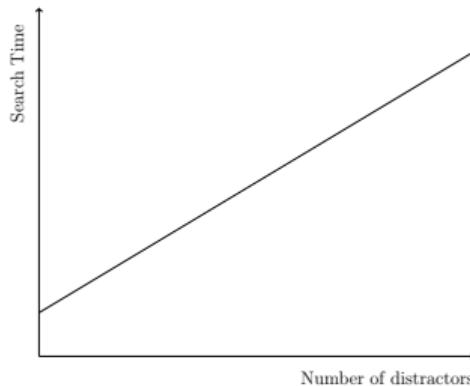
Experiment in visual Search

Find the red X in the figure



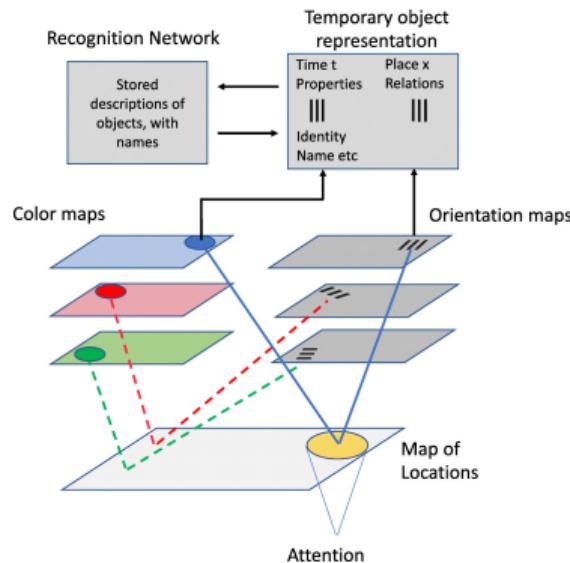
Observations

- When the target is distinguished by a single feature (color), search is almost instantaneous
- When the target is distinguished by more than one features (color and shape), search takes longer
 - It increases linearly with the number of distractors



Triesman's Feature Integration Theory (1980)

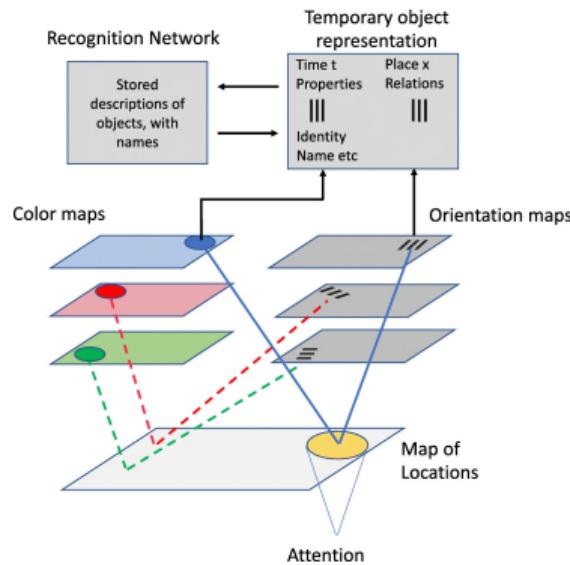
A very simple but elegant theory



- 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 are mapped
 - ▶ “Where” and **not “what”**

Triesman's Feature Integration Theory (1980)

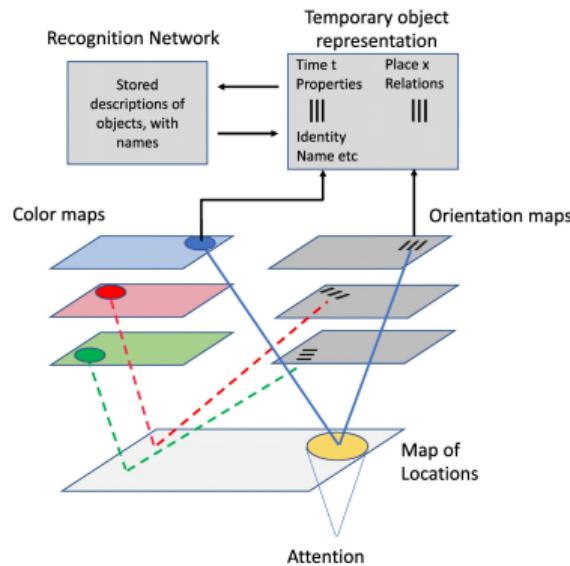
contd.



• Stage II. Attentive (late) vision

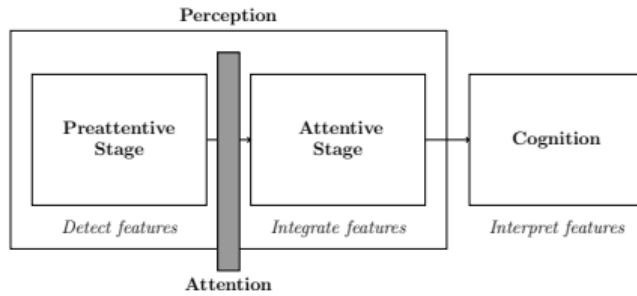
- ▶ Attention “glues” the features together
- ▶ Required for localization
- ▶ Such integrated entities came to be called “visual objects”
 - ▶ Conjunction of properties
 - ▶ Limited capacity
 - ▶ Features within same attentional focus can be encoded as belonging to the same object

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

We shall conduct an experiment instead

End of Module 04-01

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
 - ▶ Contours are fragmented
 - ▶ There are spurious edges
- Human Vision System constructs the object contours through **perceptual grouping**



Gestalt psychology / Law of Prägnanz

Seeing the whole, rather than the parts

A B C D E F G H I J K L M N O P
Q R S T U V W X Y Z à á é í ö á
b c d e f g h i j k l m n o p q r
s t u v w x y z à á é í ö á 1 2
3 4 5 6 7 8 9 0 (\$ € € . , ! ?)

"whole is greater than its parts"

“Forest before trees”



References:

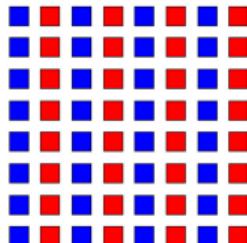
- Navon. Forest before trees: ...
- Grice, et al., Forest before trees: ... (criticism)

“Principles” of perceptual grouping

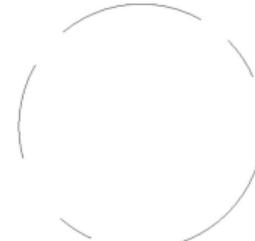
Empirical rules – Outcome of experiments by Gestalt scientists



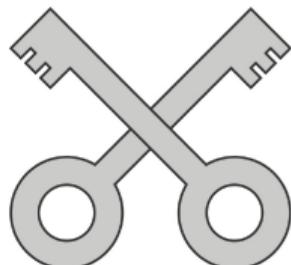
By proximity



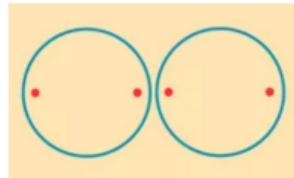
By similarity



By closure



By continuity



By common region

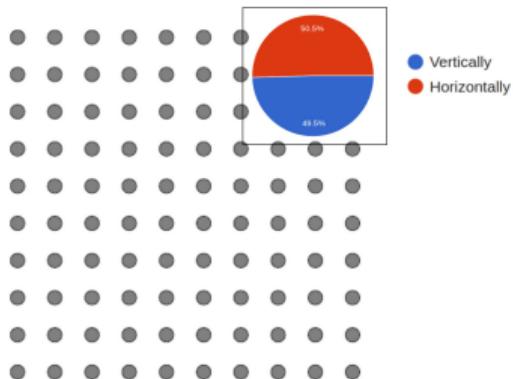


By parallelism

What is there is a conflict?

Our experiments

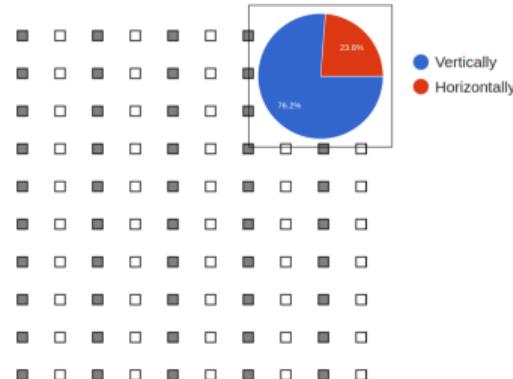
Experiment 1: Proximity – horizontal vs. vertical



Dots are equally separated horizontally and vertically.

Weak preference to horizontal grouping

Experiment 2: Similarity vs. Proximity

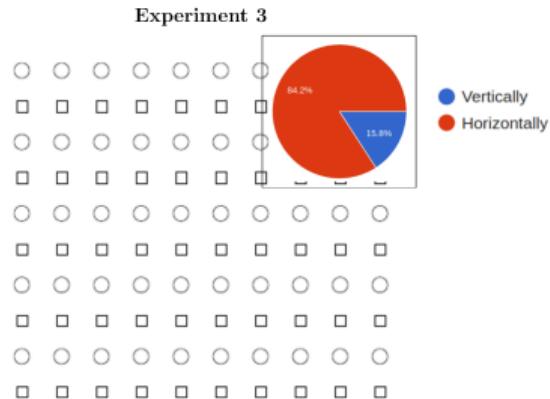


Dots are equally separated horizontally and vertically.

Strong preference for similarity than proximity

What is there is a conflict?

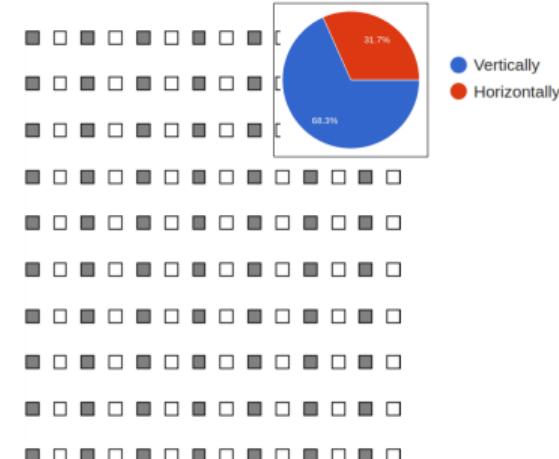
Our experiments



Dots are closer vertically than horizontally.

Strong preference for shape similarity than proximity

Experiment 4: Similarity vs. proximity



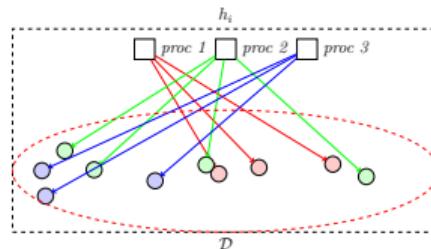
Dots are closer horizontally than vertically.

Strong preference for color similarity than proximity

Bayesian formulation for perceptual grouping

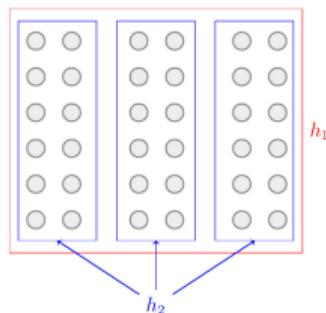
Based on the principle of Occam's razor

- Data is a set of visual elements $\mathcal{D} = \{d_j\}$.
- Data is assumed to be generated by a set of K independent processes $\mathcal{C} = \{c_k\}$
 - ▶ Each process represents a “visual object”
- Hypothesis space $\mathcal{H} = \{h_i\}$, where
 - ▶ A hypothesis h_i represents an association between the processes and data
 - ▶ Each hypothesis has a prior probability
 - ▶ The data represent “goodness of fit”
- Inference: $h^* = \text{argmax}_i P(h_i \mid \mathcal{D})$



Bayesian formulation for perceptual grouping

Illustration



- h_1 : higher prior, lower goodness of fit (sparse)
- h_2 : lower prior, higher goodness of fit (dense)

Recall discussions in taxonomy learning

Grouping by common fate

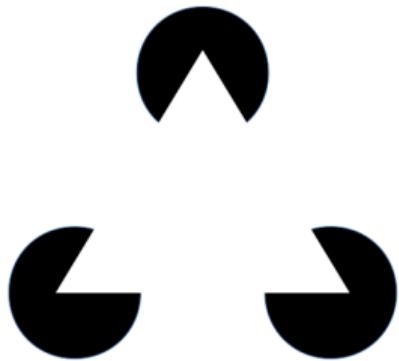
Moving together as a group



Example videos:

- Common fate
- Starlings flying

Closure and illusions



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



- Amodal completion
- The black triangle is occluded!

EdPuzzle: Application in design

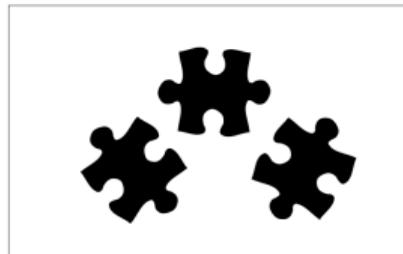
Object-ground separation

Foreground-background separation



Object-ground separation

General rules: Can be explained with Occam's razor



Closed shapes are objects



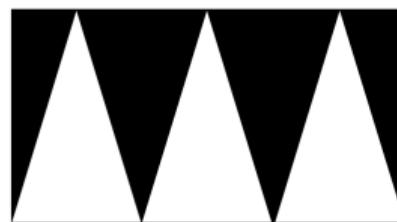
Convex shapes are objects



Symmetric shapes are objects



Shapes at bottom are objects



Shapes with fat bottom are objects



Known shapes are objects

Bistability

What do you see in the picture ?



Quiz



Quiz 04-02

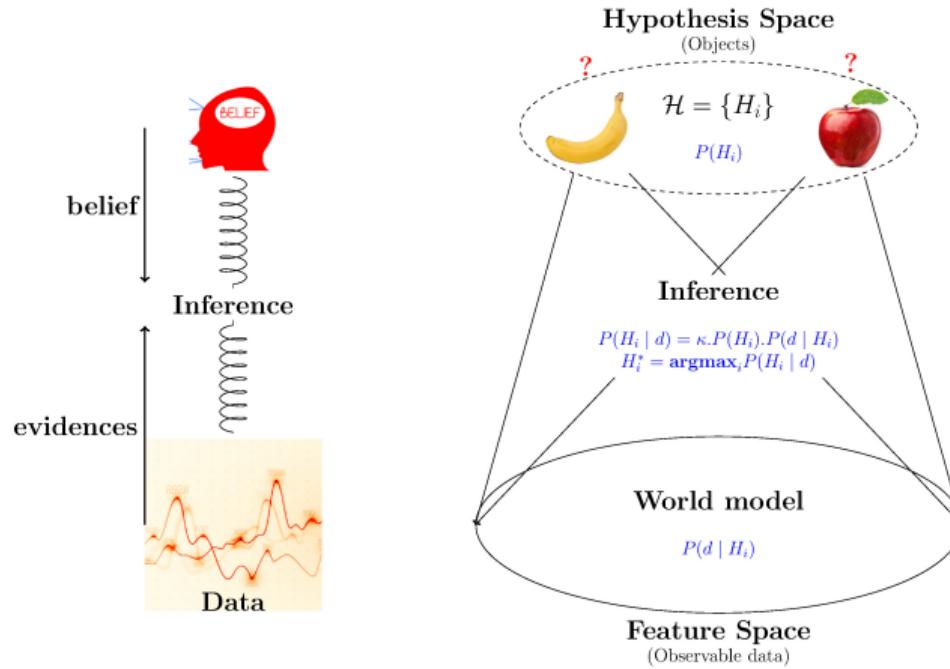
End of Module 04-02

Biological Vision and Applications

Module 04-03: Object Recognition

Hiranmay Ghosh

Bayesian Model for object recognition



Bayesian Model for object recognition

$$O^* = \operatorname{argmax}_i P(O_i | v)$$

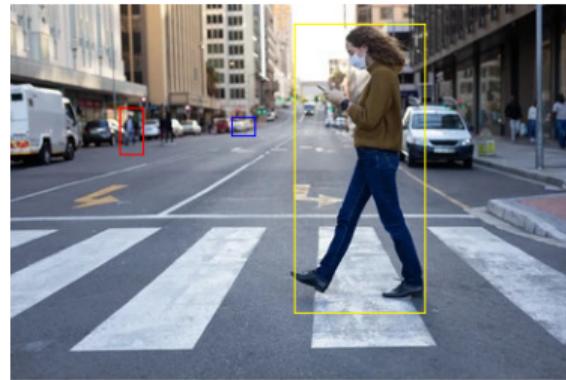
where

$$P(O_i | v) = \frac{P(O_i).P(v|O_i)}{P(v)} = k.P(O_i).P(v | O_i)$$

O_i : Object hypothesis

v : Observed 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 this object ?

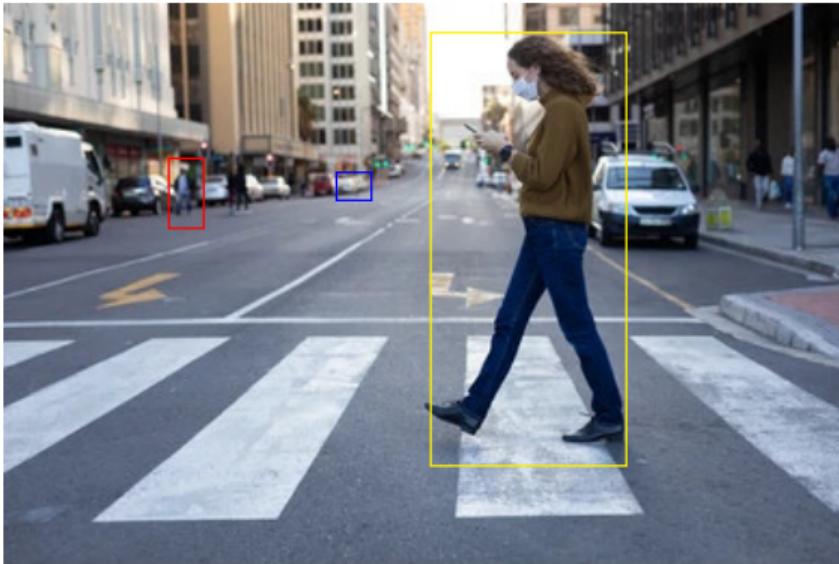


Context matters !



- Seeing the whole provides the cues for identifying the parts

Another example



Context is especially useful for robust interpretation in imperfect images

- Ambiguous features, blur, occlusion, clutter, etc.

In-context object recognition

$$P(O_i | v) = k \cdot P(O_i) \cdot P(v | O_i), \quad v = (v_l, v_c)$$

$$\begin{aligned} P(O_i | v_l, v_c) &= P((O_i | v_c) | v_l) = \frac{P(O_i | v_c) \cdot P(v_l | O_i, v_c)}{P(v_l)} \\ &= \kappa \cdot P(O_i | v_c) \cdot P(v_l | O_i, v_c) \end{aligned}$$

$P(O_i | v_c)$: Prob of O_i to appear in a specific context v_c

$P(v_l | O_i, v_c)$: Visual model for O_i in the same context v_c

We can assume, visual features of an object is independent of context (**how?**)

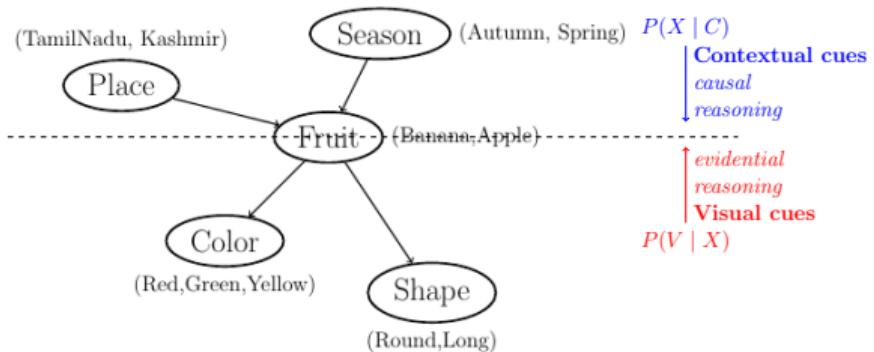
$$P(v_l | O_i, v_c) = P(v_l | O_i)$$

$$P(O_i | v_l, v_c) = \kappa \cdot P(O_i | v_c) \cdot P(v_l | O_i)$$

Example of context based reasoning

Context can be of different kinds – maybe external to the image

$$P(X \mid C, V) = \kappa \cdot P(X \mid C) \cdot P(V \mid X)$$



How do you decide if a photo is taken in the day or in the night ?



Visual context

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

- O_i is the manifestation of an object instance in a certain location of a scene
 - ▶ ... not just an object class
- Let $O_i = (o_i, x_i, \sigma_i)$ where
 - ▶ o_i : object class
 - ▶ x_i : location in image
 - ▶ σ_i : appearance (scale, orientation, etc.)
- $P(O_i | v_c)$ represents an object of a class to appear in a specific location in an image with a certain appearance

$$P(O_i | v_c) = P(o_i, x_i, \sigma_i | v_c) = P(\sigma_i | o_i, x_i, v_c).P(x_i | o_i, v_c).P(o_i | v_c)$$

In-context object recognition

Significance of the decomposition



- $P(o_i | v_c)$: Probability of an object class to appear in a context
- $P(x_i | o_i, v_c)$: Probability of the location where an object class appears in a context
- $P(\sigma_i | o_i, x_i, v_c)$: Probability of the appearance of an object class when it appears in a certain location in an image

$$P(O_i | v_l, v_c) = \underbrace{\kappa.P(\sigma_i | o_i, x_i, v_c).P(x_i | o_i, v_c).P(o_i | v_c)}_{\text{context}}. \underbrace{P(v_l | O_i)}_{\text{evidence}}$$

How do we characterize a context ?

What is v_c ?



- Plate is recognized by it's context
- Other objects in the scene creates the context
 - ▶ $v_c = \{ \text{Fork, knife, table-mat} \}$
- How do you recognize those objects?
 - ▶ A chicken-and-egg problem?

Recap: Forest before the trees ?



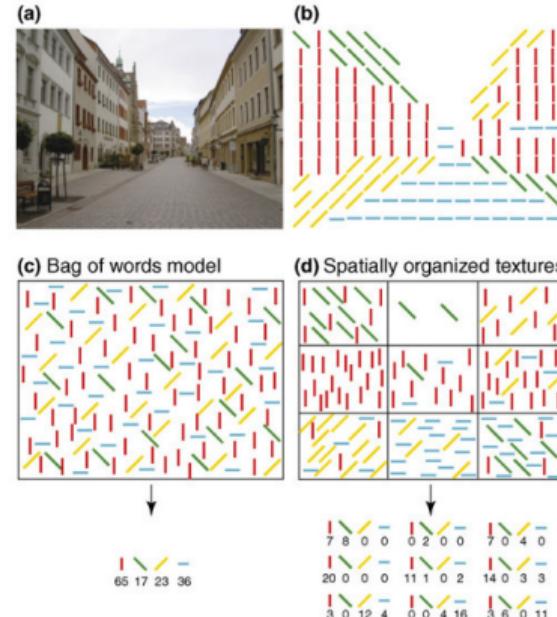
Do the “scenes” have some distinctive features?



Spatial envelop representation

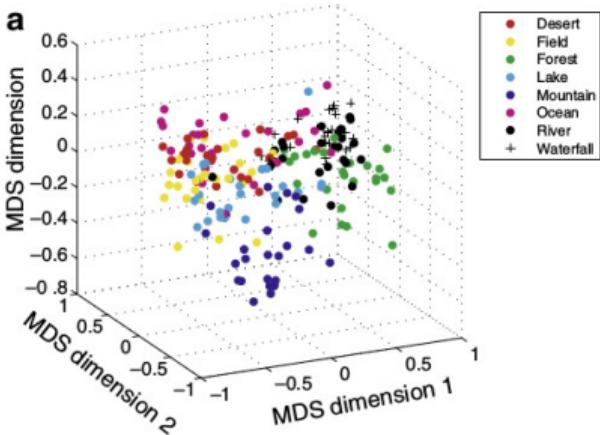
A holistic representation of a scene layout

- The edges in a scene constitutes a definite pattern
 - ▶ Statistical pattern characterizes a scene
- Recall natural scene statistics
 - ▶ Happens in early (pre-attentive) vision
- Computational Model:
 - ▶ Global and local statistics
 - ▶ Abstract image features



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

Distinguishing scene classes with spatial envelop representation



The three MDS axes represent three abstract features of an image: openness, ruggedness and expansion .

Quiz



Quiz 04-03

End of Module 04-03

Biological Vision and Applications

Module 05-01: Visual attention

Hiranmay Ghosh

Attention

Coping up with huge data volume

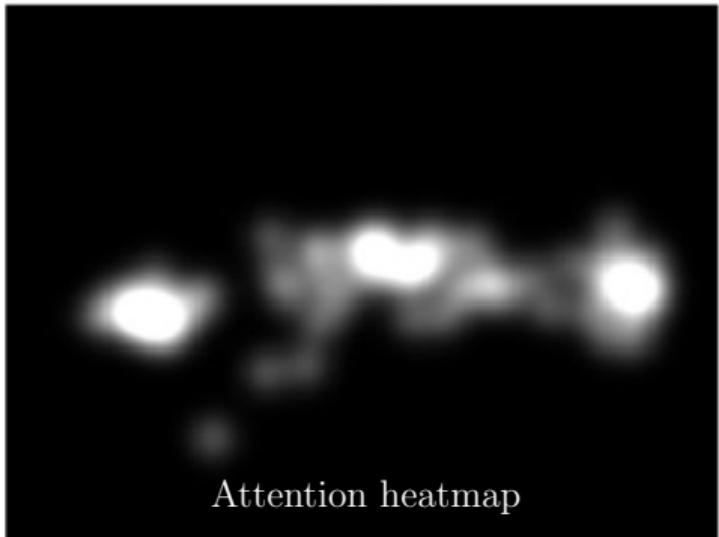
- About 1 mn optic nerves come out of each eye
- Data rate is 10 Mbps, assuming
 - ▶ 1 bit of data for each nerve
 - ▶ Refresh rate: $\frac{1}{10}$ th sec



What is the data rate of a video camera of modest resolution 1280×768 operating at 30fps?

Attention

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

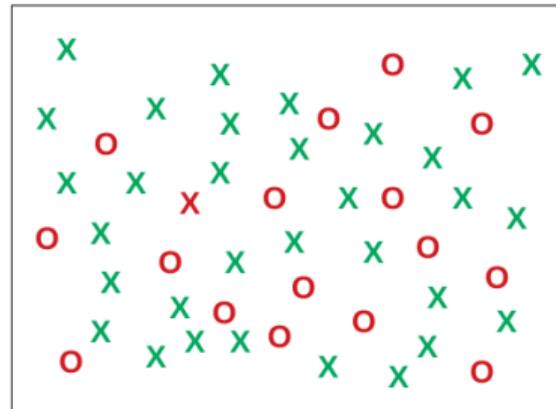


Attention heatmap

Attention leads to change blindness (Video illustration)

- Focus on the center of the disk and see other objects disappearing

Bottom-up and top-down attention



Bottom-up attention – stimulus driven (exogenous)
– spontaneous

Top-down attention – task driven (endogenous)
– knowledge-based

Bottom-up and top-down attention

More examples



Bottom-up attention

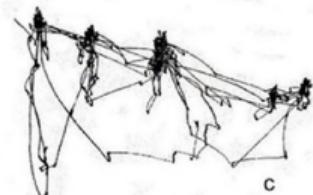
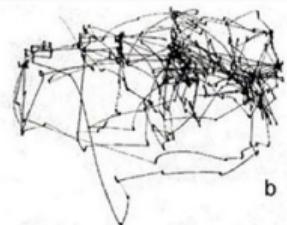


Where is my cat ?

Top-down attention

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
 - ▶ ...

Yarbus experiments ... and more

Modeling attention

- Classical approaches
 - ▶ Image feature based
 - ▶ Surprise based
 - ▶ Object based
 - ▶ Context based
- Neural network based approaches
 - ▶ We reserve for the future
- Do bottom-up and top-down attention work together ?

Quiz



Quiz 05-01

End of Module 05-01

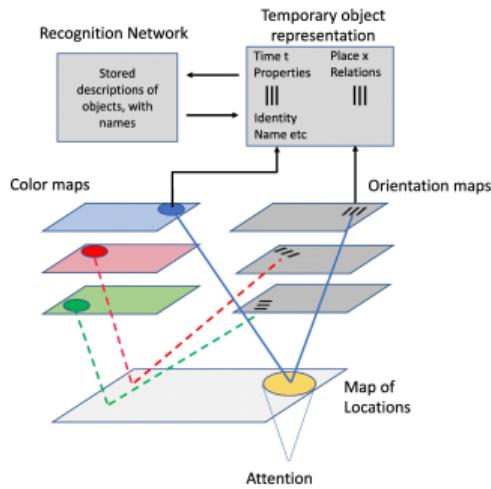
Biological Vision and Applications

Module 05-02: Cognitive attention models

Hiranmay Ghosh

Cognitive Models

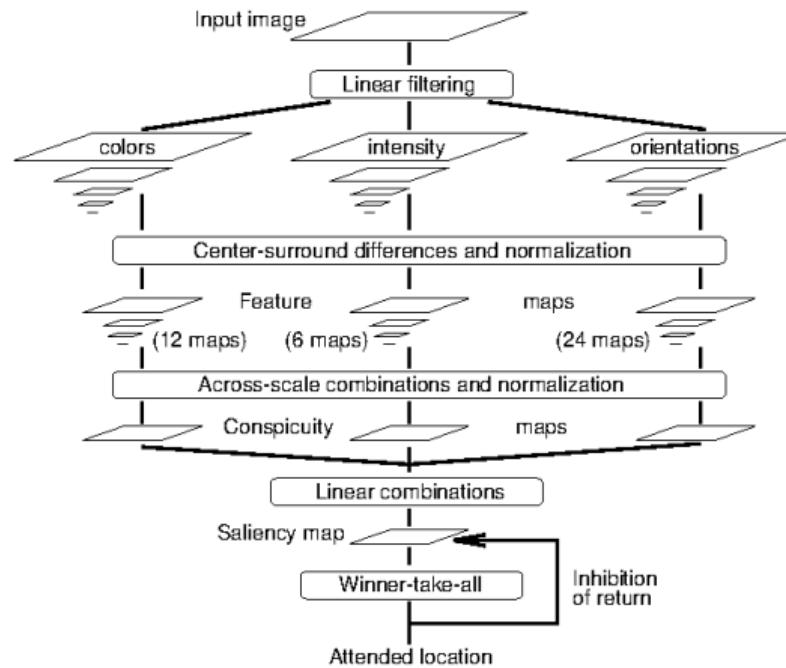
Motivated by Feature Integration Theory



- Based on the observations
 - ▶ Early vision distinguishes local contrasts
 - ▶ ... colors, edges
 - ▶ Features are subsequently integrated
 - ▶ Treisman's Feature Integration Theory
 - ▶ Higher acuity at central vision (5°)
 - ▶ ... lower at paracentral / macular ($8 - 18^\circ$)

Itti's model (1998)

Overview

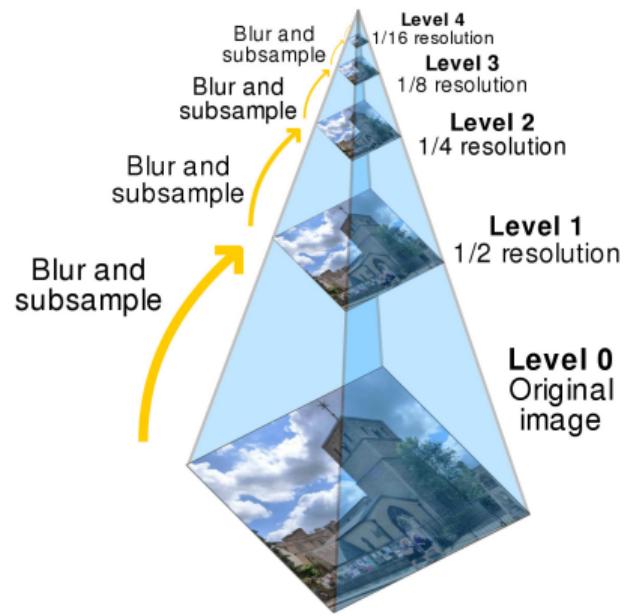
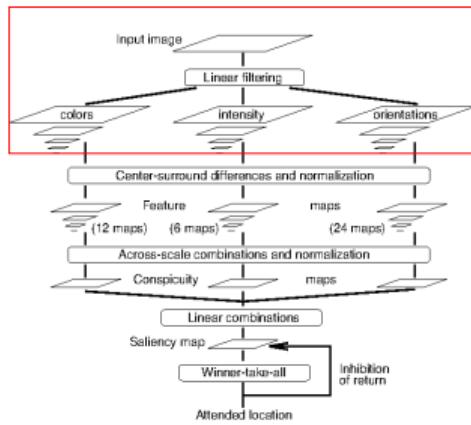


Itti, et al., A saliency based search mechanism ...

Itti's model: Stage 1

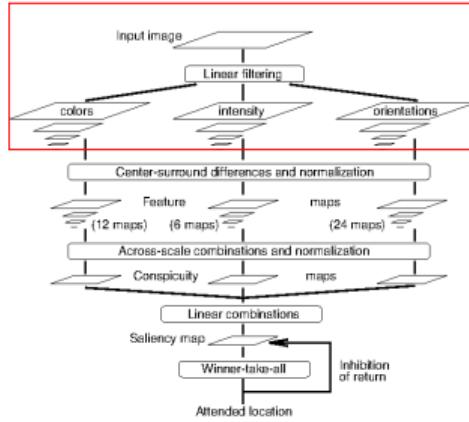
Multi-resolution image analysis

- Multi-resolution analysis of input image
 - ▶ Using Gaussian pyramids (9 scales: 0 – 8)



Itti's model: Stage 1

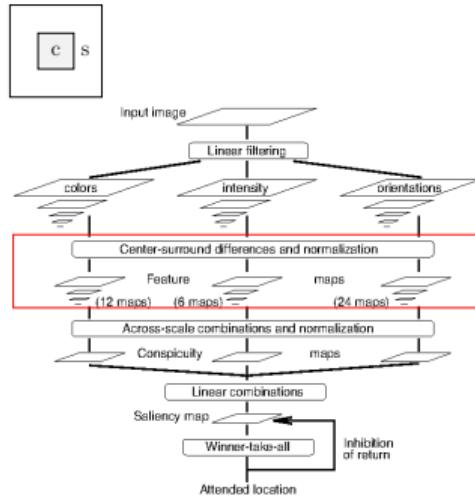
Feature extraction



- 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

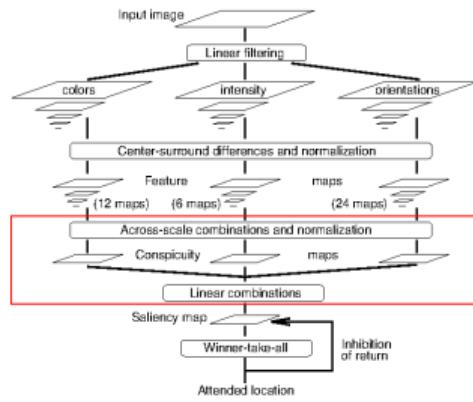
Center-surround operations: Multi-scale feature maps



- Center-surround difference computed for each of 7 features for every location
- 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
 - ▶ $\mathcal{F} = | F(c) \ominus F(s) |$
- 6 scales for each feature
- $7 \times 6 = 54$ “feature maps” (contrasts)
 - ▶ Each represents local contrast at a location based on a feature at a certain scale

Itti's model: Stage 3

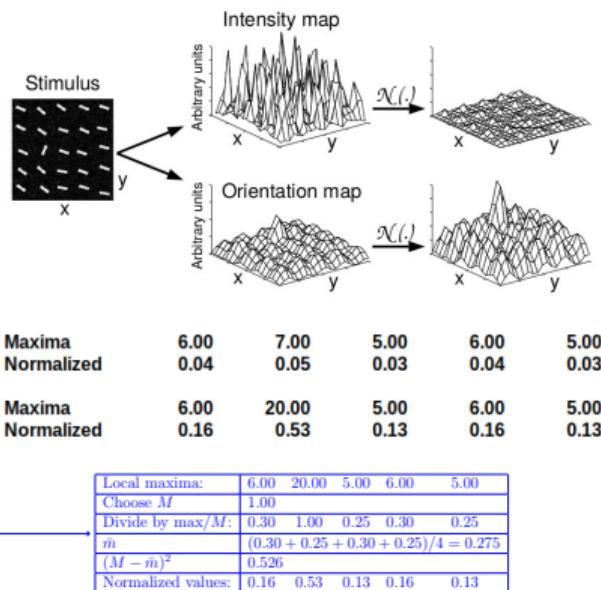
Combining the features: Conspicuity and Saliency Maps



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

Itti's model: Stage 3

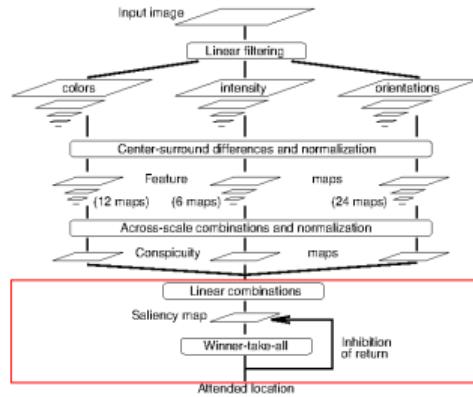
Normalization



- Two reasons to normalize
 - ▶ Features are at arbitrary scale
 - ▶ Normalize to a fixed range $[0, M]$
- Some feature may have many nearly equal peaks, indicating texture
- Steps:
 - ▶ Choose M
 - ▶ Normalize so that the global max = M
 - ▶ Compute the average of all other local maxima \bar{m}
 - ▶ Multiply the map by $(M - \bar{m})^2$

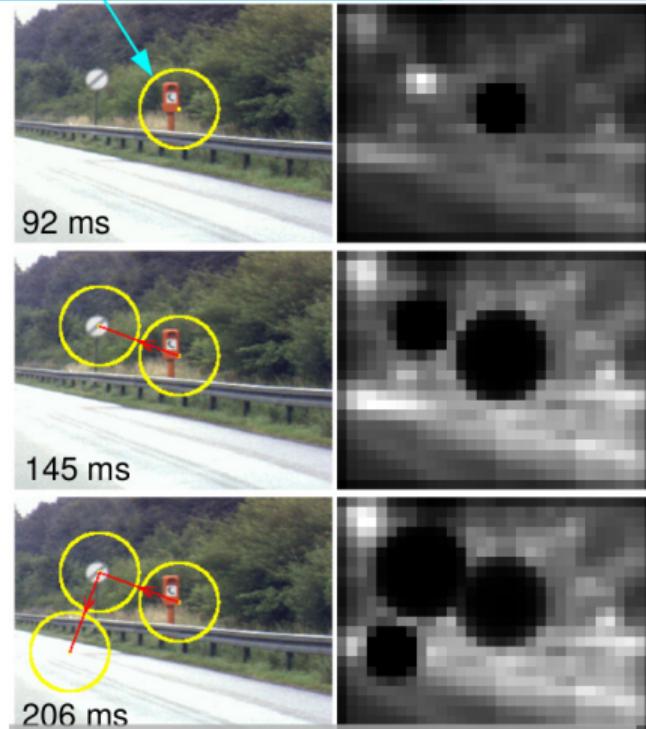
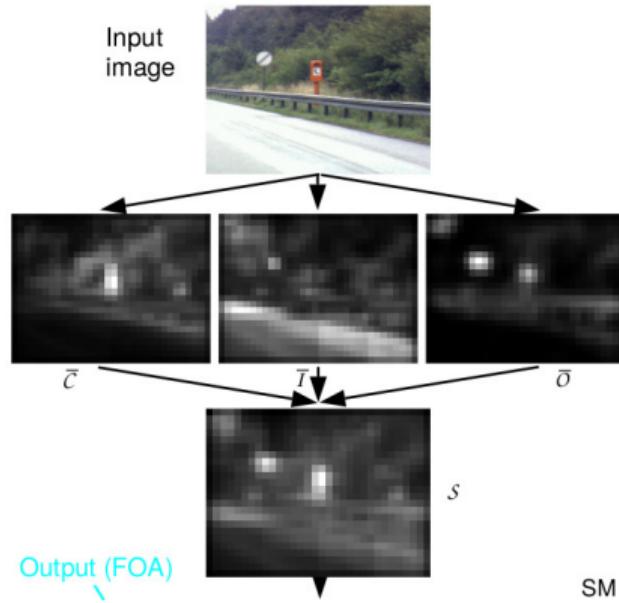
Itti's model: Stage 4

"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 tire out.
 - ▶ Attention moves to the location with next highest salience.

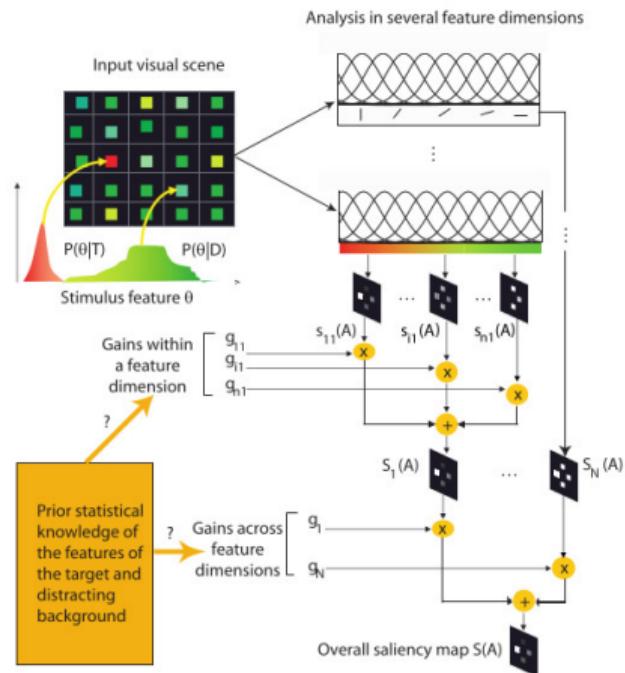
Sample Results



Discussions

- Remains a reference model till date
 - ▶ WTA and RI policies are common to all classical models
- 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

Adaptation to top-down attention

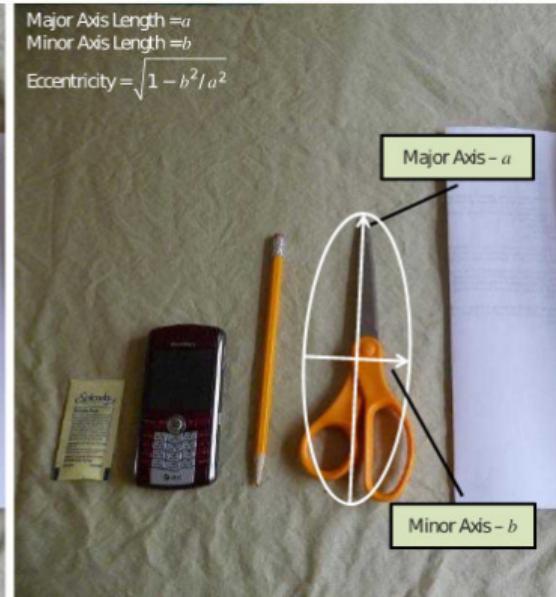
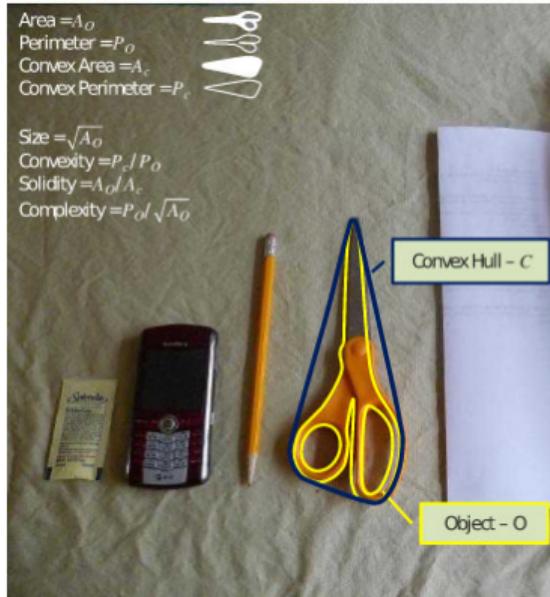


- Visual search task
- Weights assigned to features based on task requirement
- Weights learned from statistical features of target and distractors
- Inflexible

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
 - ▶ Sequence?
 - ▶ Color first. No consensus of other features

Khan, et al. Top down color attention ...

Quiz



Quiz 05-02

End of Module 05-02

Biological Vision and Applications

Module 05-03: Surprise based attention models



Hiranmay Ghosh

Surprise based model

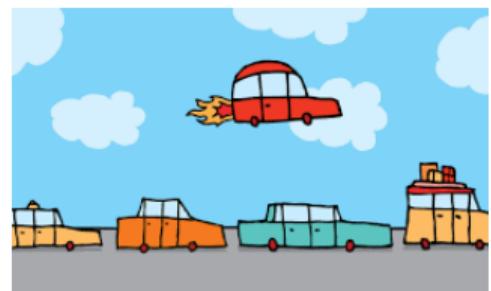
- Information-theoretic model
- Probabilistic (Bayesian) model

Information theoretic model

- Info theoretic model is based on Shannon's information theory:

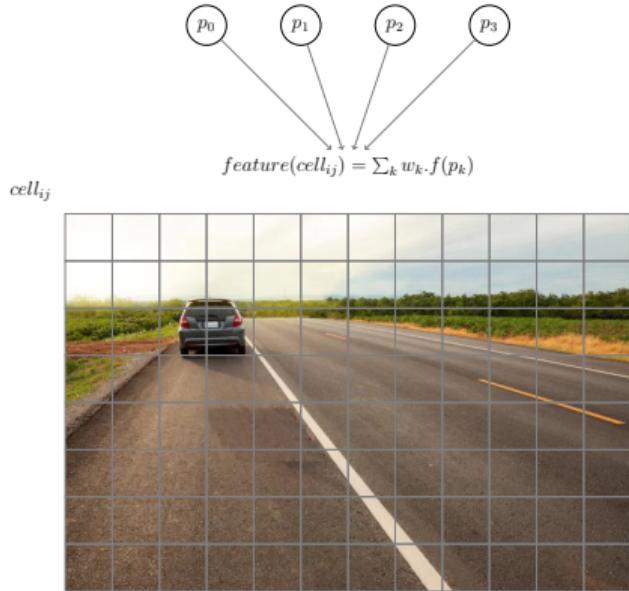
- ▶ The event that is least likely to occur has the maximum information value
- ▶ Self-information of an event x : $-\log P(x)$

- 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

A generative model of an image

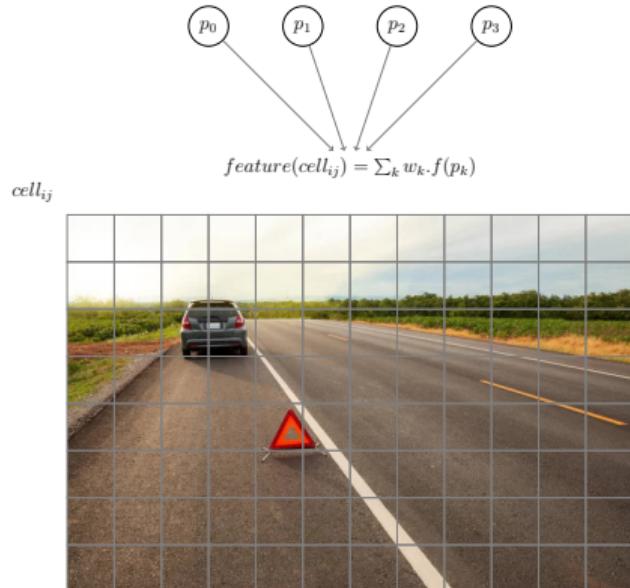


- An image region is a manifestation of some underlying hidden processes
 - ▶ A weighted sum of contributions from each process
 - ▶ The processes are unknown, hidden, independent of each other
 - ▶ The weights are **not** à-priori known
- Use Natural Scene Statistics
 - ▶ Learn the processes and the weights from many observations (different types of scenes)

Independent Component Analysis

A generative model of an image

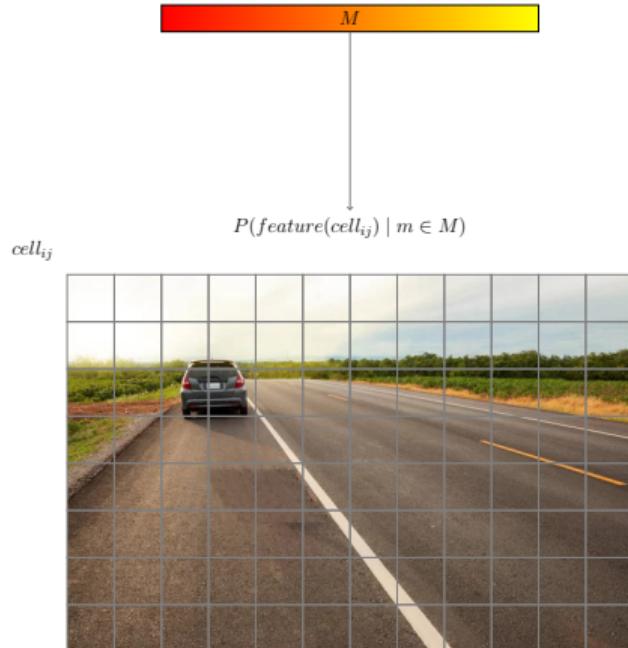
contd ... What is unexpected?



- Model a new scene with the learned features & weights
 - ▶ Select weights for the processes for best fit for the overall image
- There will be some outlier regions, which do not fit
 - ▶ Have least probability to occur (most informative)
- These image regions are the salient ones

Bayesian model

Based on “surprise” – brings in experiential factor

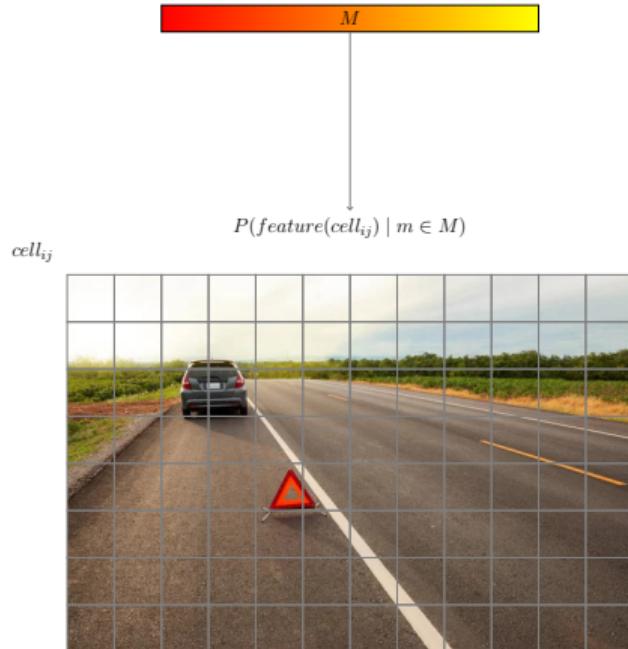


- D : observed data (features in a cell)
- M : a continuous range of states
- $p(m)$: the prior pdf for states
- $p(m \mid D)$: the posterior pdf for the states
 - ▶ after experiencing some data D
- The surprise factor of the data D :
 - ▶ Change in pdf of M as a result of observing D

$$\begin{aligned} S(D) &= \text{KLD}(p(m), p(m \mid D)) \\ &= \int_m p(m) \cdot \log \frac{p(m)}{p(m \mid D)} \cdot dm \end{aligned}$$

Surprise ... Bayesian Model

... contd.



Surprise factor for data D :

$$S(D) = \int_m p(m) \cdot \log \frac{p(m)}{p(m|D)} \cdot dm$$

Baye's Theorem:

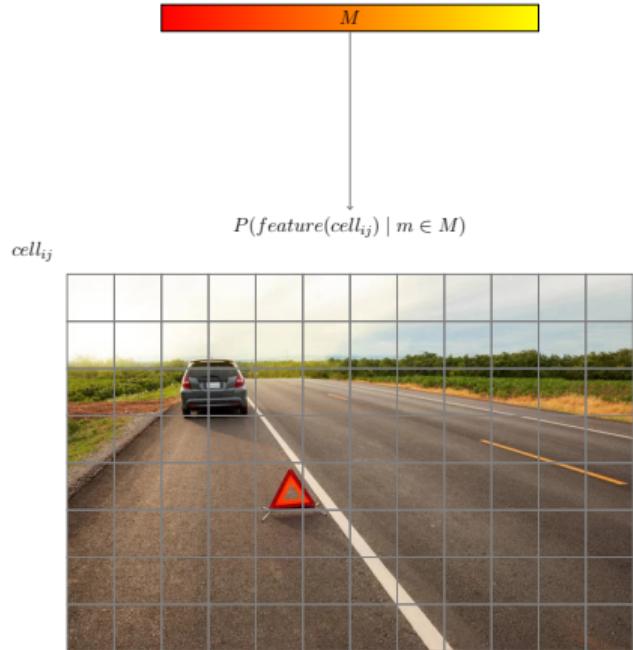
$$P(m | D) = \frac{p(m) \cdot P(D|m)}{P(D)}$$

Using Baye's theorem and simplifying:

$$S(D) = \log P(D) - \int_m p(m) \cdot \log P(D | m) \cdot dm$$

Surprise ... Bayesian Model

... incremental update and change awareness



- Let data $D = D_1, D_2, \dots$ (a time series)
- After observing D_1 :
 - ▶ $p_1(m) = p(m \mid D_1) = \frac{p(m).P(D_1|m)}{P(D)}$
 - ▶ This serves as the prior for next observation D_2
- The model of the environment is incrementally built
- Leads to change awareness

Eye movement ?

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

Quiz



Quiz 05-03

End of Module 05-03

Biological Vision and Applications

Module 05-04: Context-based attention models



Hiranmay Ghosh

Context-based Model

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

$$P(O \mid v_l, v_c) = \frac{1}{P(v_l \mid v_c)} \cdot (v_l \mid O, v_c) \cdot P(O \mid v_c)$$

Substituting $O = (o, x, \sigma)$ and expanding:

$$P(O \mid v_c) = P(\sigma \mid x, o, v_c) \cdot P(x \mid o, v_c) \cdot P(o \mid v_c)$$

Conditional independence: $P(v_l \mid O, v_c) = P(v_l \mid O)$

Substituting and rearranging:

$$P(O \mid v_l, v_c) = \frac{1}{P(v_l \mid v_c)} \cdot \underbrace{P(v_l \mid O)}_{\textcircled{1}} \cdot \underbrace{P(\sigma \mid x, o, v_c)}_{\textcircled{3}} \cdot \underbrace{P(x \mid o, v_c)}_{\textcircled{2}} \cdot P(o \mid v_c)$$

1. Bottom-up saliency (task-independent)

What catches spontaneous attention

- $1/P(v_l | v_c)$:

- ▶ How unlikely is v_l given a context v_c
 - ▶ Independent of object hypothesis O
 - ▶ surprise factor
- ▶ Represents bottom-up saliency
- ▶ Bayesian model brings in experiential factor (temporal)

2. Top-down saliency (task specific)

Where to look for

- $P(x | o, v_c).P(o | v_c)$
 - ▶ $P(o | v_c)$: Prob of an object class to appear in a context
 - ▶ $P(x | o, v_c)$: Prob of an object class to appear at a certain location in a context
 - ▶ ... given that it appears
- $P(x | o, v_c).P(o | v_c) = P(o, x | v_c)$:
 - ▶ Represents the task-specific context-driven saliency of location

3. Appearance model (task specific)

What to look for

- $\underline{P(v_I \mid O).P(\sigma \mid x, o, v_c)}$
 - ▶ $P(\sigma \mid x, o, v_c)$: appearance (scale, orientation) of the object
 - ▶ ... when it appears at a certain location
 - ▶ $P(v_I \mid O) = P(v_I \mid o, x, \sigma)$: The expected visual features
 - ▶ ... when the object appears at a certain location with a certain appearance

In Summary

$$P(O \mid v_l, v_c) = \frac{1}{P(v_l \mid v_c)} \cdot P(v_l \mid O) \cdot P(\sigma \mid x, o, v_c) \cdot P(x \mid o, v_c) \cdot P(o \mid v_c)$$

1

3

2

1. Bottom-up saliency (s_b): $\frac{1}{P(v_l \mid v_c)}$

► What spontaneously draws attention

2. Top-down saliency (s_t): $P(o, x \mid v_c) = P(x \mid o, v_c) \cdot P(o \mid v_c)$

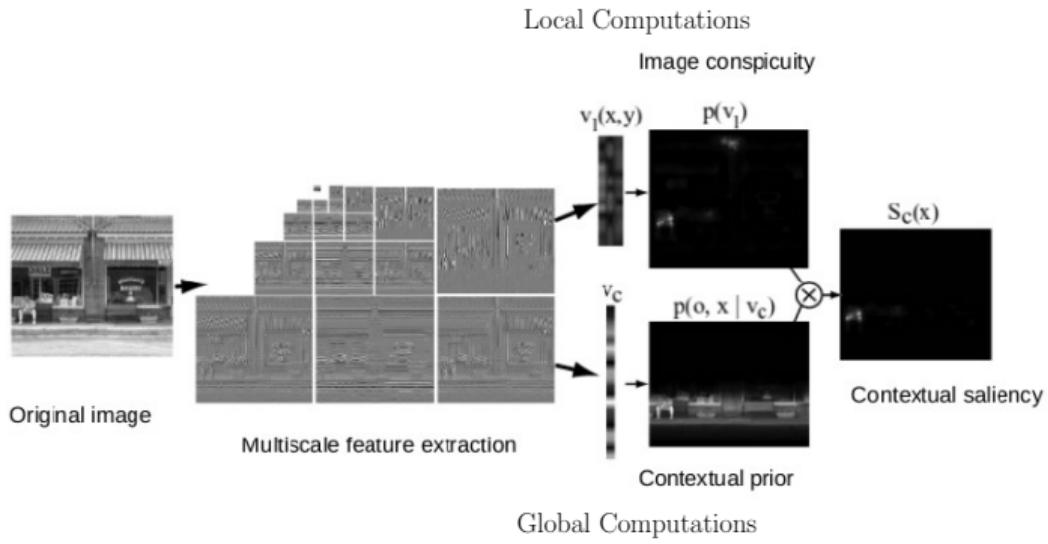
► Where to look for for an object class

3. Visual features (v): $P(v_l \mid o, x, \sigma) \cdot P(\sigma \mid x, o, v_c)$

► What to look for at a certain location to detect an object

- Overall saliency (where): $s_c = s_b \times s_t$

- Feature to look for (what): v



Quiz



Quiz 05-04

End of Module 05-04

Biological Vision and Applications

Module 05-05: Attention: Miscellaneous topics

Hiranmay Ghosh

Evaluation

Dynamic: Compare predicted gaze movement with actual human eye movement



Human fixations:

1 – 2 – 4 – 5 – 3 – 6

Predicted fixations:

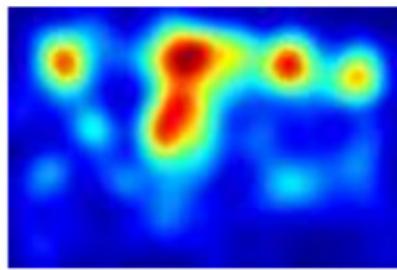
1 – 3 – 6 – 2 – 4 – 5

What is the distance between the two sequences ?

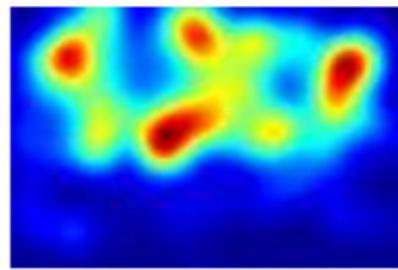
Levenshtine Distance

Evaluation

Static: Compare predicted attention heatmap with actual human attention heatmap



Experimental Heatmap
 $H(x, y)$



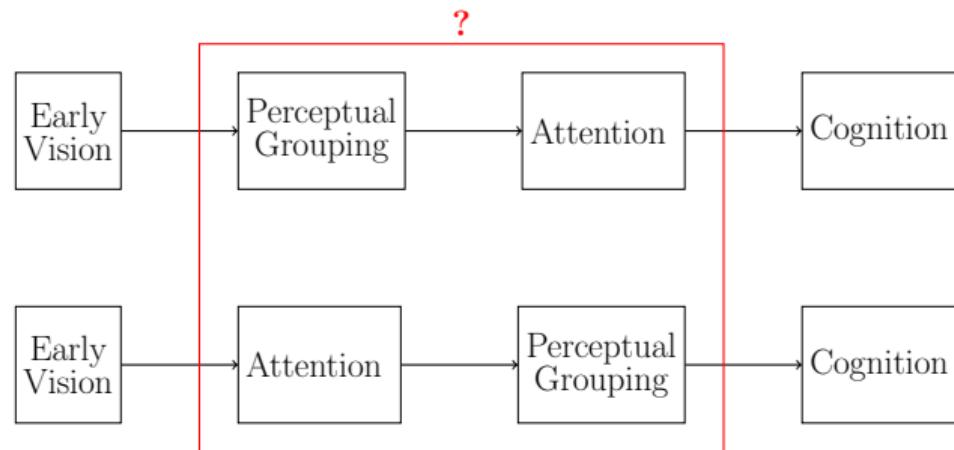
Predicted saliency
 $S(x, y)$

How are the values $H(x, y)$ and $S(x, y)$ correlated ?

Pearson Linear Correlation Coefficient (PLCC)

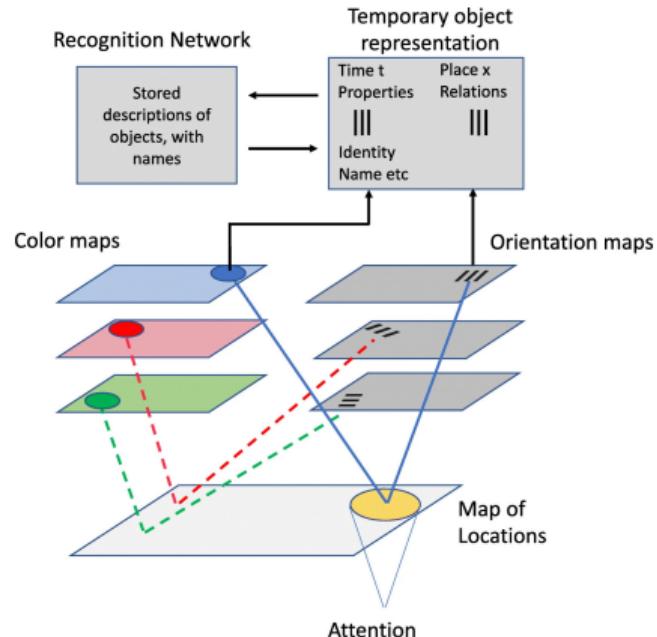
Which one first? Perceptual Grouping or Attention

Is perceptual grouping a pre-attentive or post-attentive phenomenon ?



Is attention needed for perceptual grouping ?

Recap: Treisman's feature integration theory



Suggests that attention is needed for perceptual grouping

Is attention needed for perceptual grouping ?

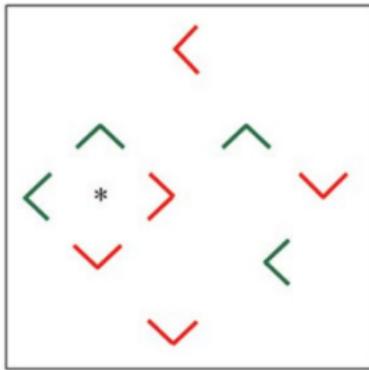
Experiment

	o	+	o	+	o	+	o	+	o	+
	o	+	o	+	o	+	o	+	o	+
	o	+	o	+	o	+	o	+	o	+
	o	+	o	+	o	+	o	+	o	+
o	+	o			o	+	o	+	o	+
o	+	o			o	+	o	+	o	+
o	+	o			o	+	o	+	o	+
					o	+	o	+	o	+
					o	+	o	+	o	+
					o	+	o	+	o	+

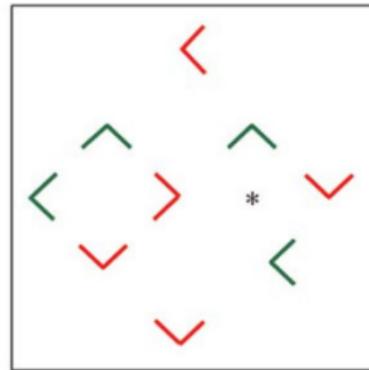
Attention is needed for perceptual grouping (similarity, proximity, etc.)

Does Perceptual Grouping affect Attention ?

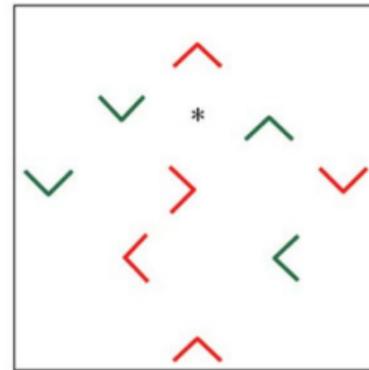
a. Inside-object



b. Outside-object



c. No-object

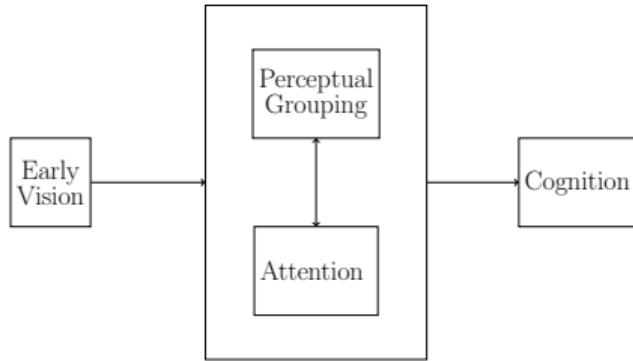


Perceptual grouping (by closure) affects attention

Objects are formed with perceptual grouping; objects draw attention

Perceptual Grouping vs. Attention

Summary



Perceptual Grouping and Attention complement each other.

Some form of perceptual grouping happens pre-attentively, some needs attention.

EdPuzzle: Visual attention (advanced topics)

Quiz



No quiz for module 05-05

End of Module 05-05

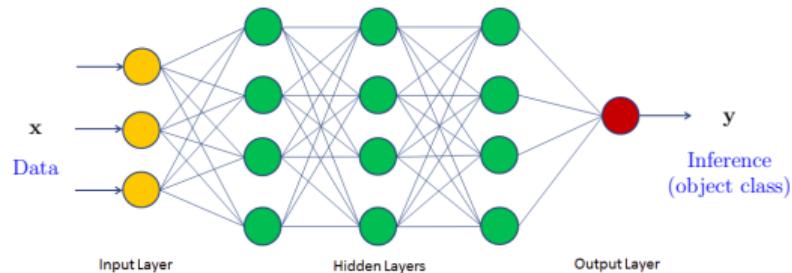
Biological Vision and Applications

Module 06-01: Introduction to Neural Networks



Hiranmay Ghosh

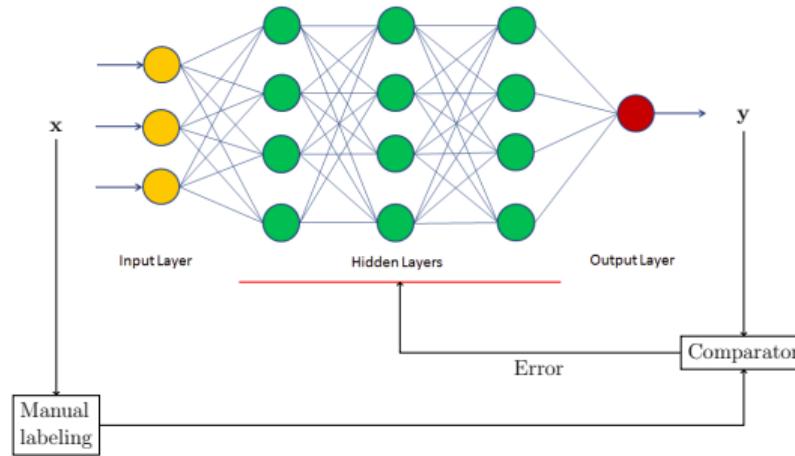
Neural Networks



- Transfer function: $y = \mathbf{W} \cdot x$
- Feed-forward network: back-propagation algorithm for training
- \mathbf{W} is a constant: deterministic output

Neural Networks

Training



- Back-propagation algorithm
 - ▶ Adjust network parameters based on error
 - ▶ Minimize error over many observations

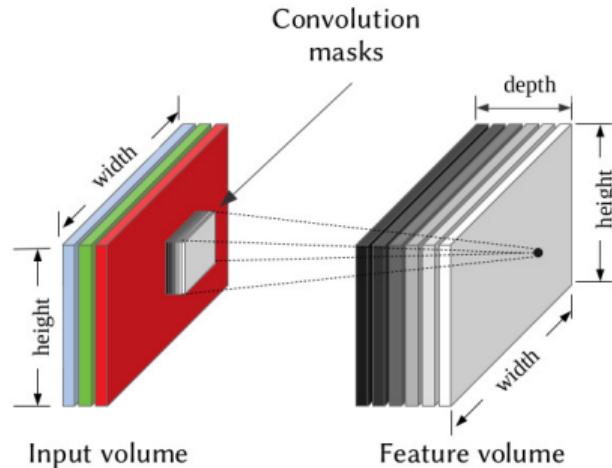
Challenges for using neural network for image processing

Training

- For a 640×480 color image
 - ▶ Number of input nodes = 927,360
 - ▶ Large number of parameters to be learned
- Early vision:
 - ▶ Image is organized in 2D
 - ▶ All image locations are to be similarly processed (contrast detection)

Convolutional Neural Networks (CNN)

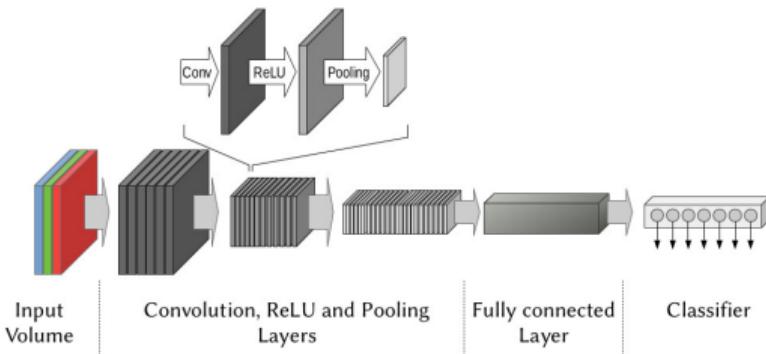
Exploits properties of early vision



- 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

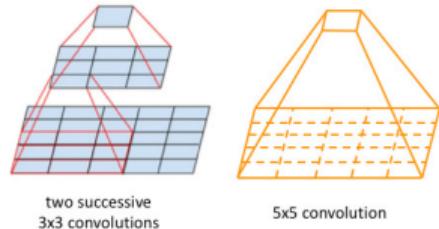
Convolutional Neural Networks

Structure

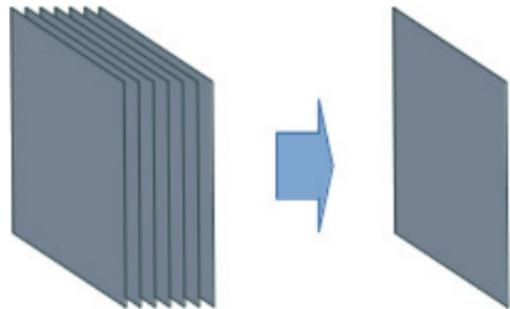


- Architecture motivated by early vision
 - ▶ **Convolution:** Aggregates information from receptive field
 - ▶ **Filtering (ReLU):** Non-linear transformation
 - ▶ **Pooling (avg / max):** Reduces information volume

On filter sizes



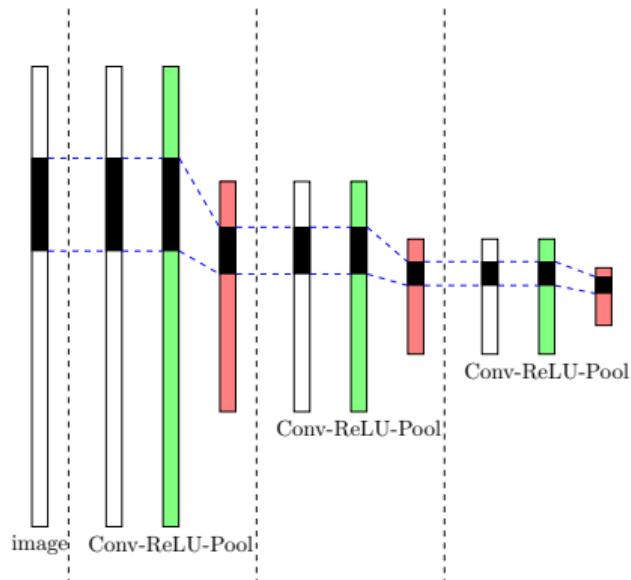
- A bank of two 3×3 filters in succession has a receptive field of 5×5
 - ▶ Can implement identical transfer function
- Which one would you prefer?



- Filter size = 1
- For “flattening” the layers
 - ▶ $y(i,j) = \sum_k w_k \cdot x_k(i,j)$

Progressive abstraction

Use of context



- 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

Some notable CNN implementations (2012 – 2015)

These implementations are reused in different contexts

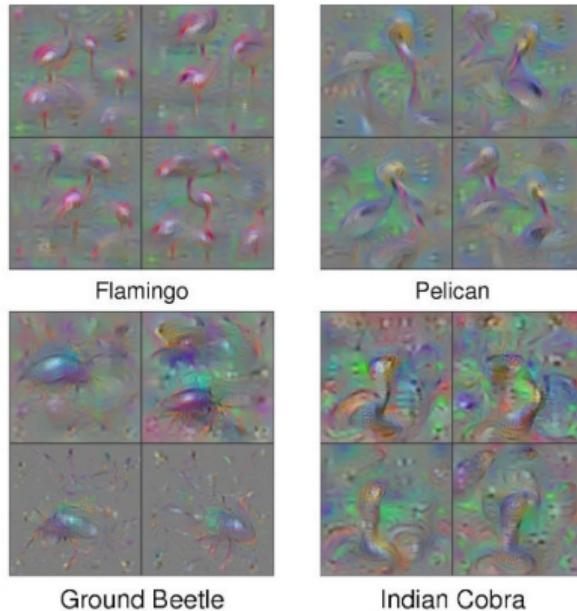
- AlexNet
- VGG
- ResNet
- GoogleNet

A feed-forward network does not learn from (runtime) experience

[Architecture comparisons \(blog\)](#)

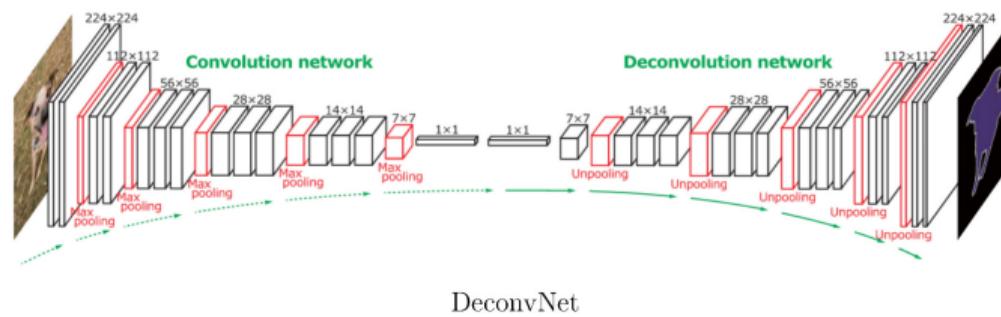
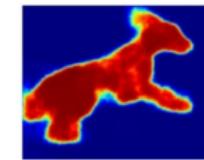
Does CNN really do progressive abstraction ?

Visualization at the last layer – just before classification



Fully Convolutional Neural Network (FCNN)

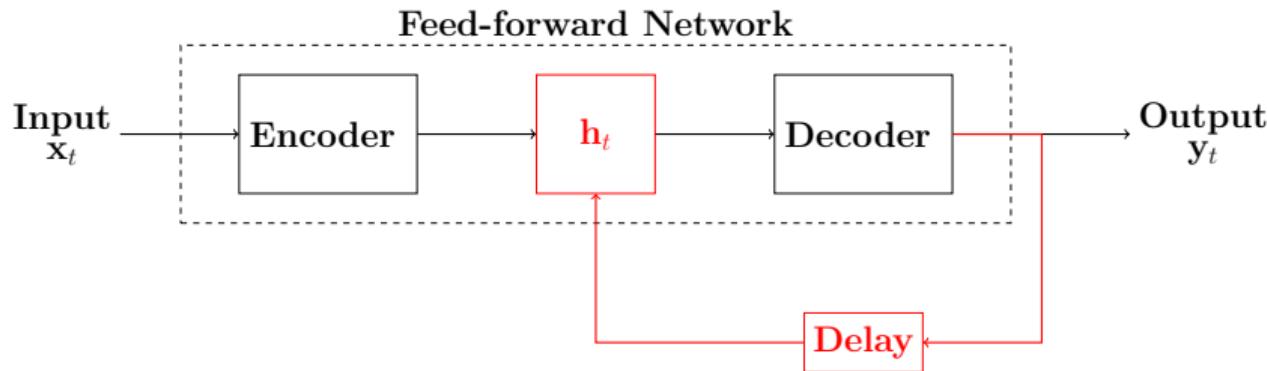
Used for Image Segmentation



Understanding DeconvNet

Recurrent Neural Network (RNN)

Tool for sequence processing tasks (natural language, video, ...)

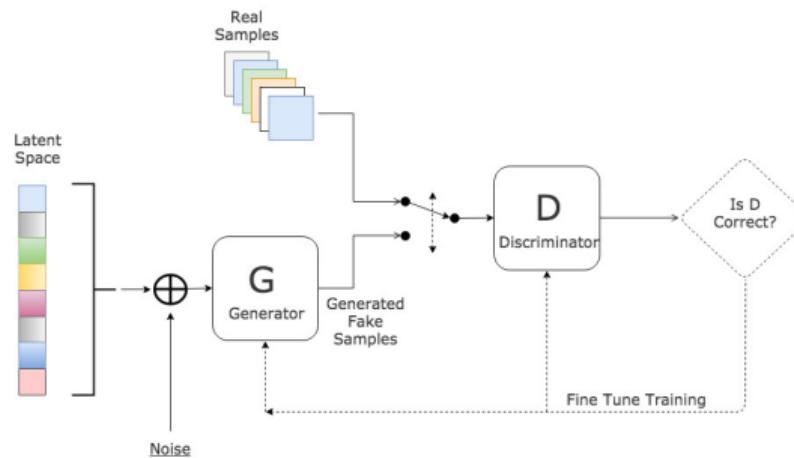


- RNN incorporates a feedback loop (with delay)
- Transfer function
 - ▶ $h_t = f(W_1 \cdot x_t + W_2 \cdot y_{t-1})$
 - ▶ $y_t = g(W_3 \cdot h_t)$
 - ▶ h_t accumulates experience

Introduction to RNN

Generative Adversarial Network (GAN)

Technology behind DeepFake, etc.



- Generator attempts to create realistic samples
- Discriminator tries to differentiate between real samples and the fakes (generated)
- Both networks are trained together – both improves with training

[GAN tutorial](#)

Quiz



Quiz 06-01

End of Module 06-01

Biological Vision and Applications

Module 06-02: Neural Network based attention models

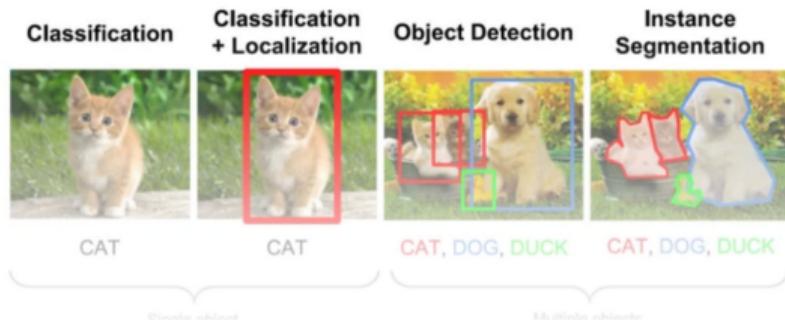


Hiranmay Ghosh

Classification-localization-segmentation

Role of attention

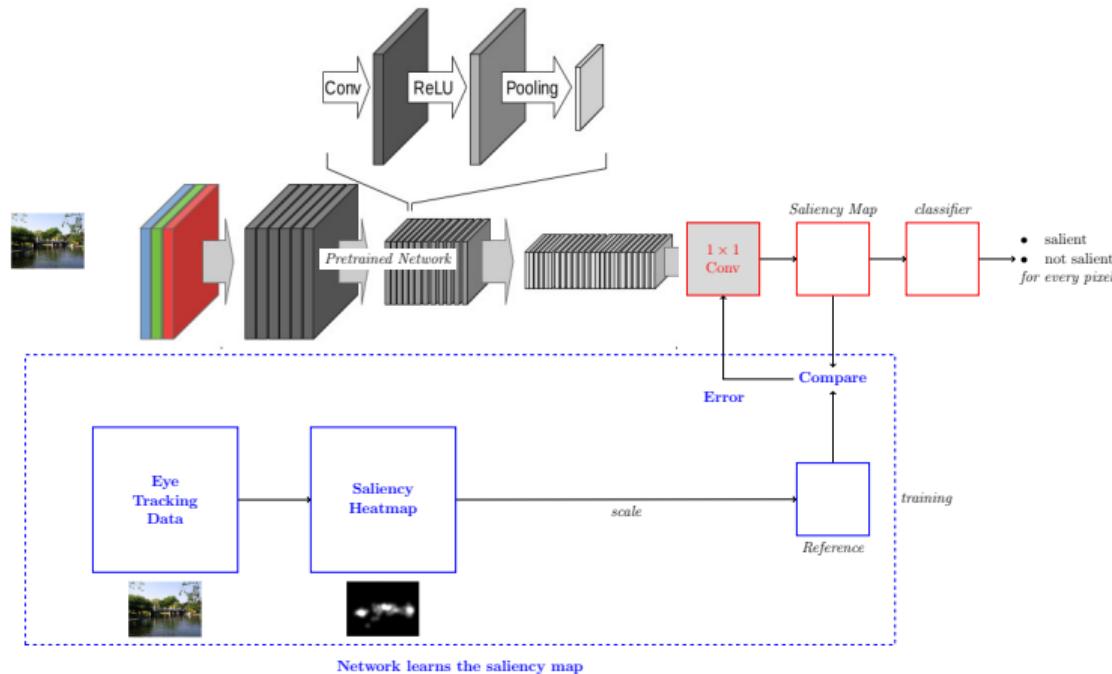
Progressive development of CNN based image processing



Attention and object recognition goes hand-in-hand

- The objects determine focus of attention
- Object recognition / segmentation takes place where there is attention

Basic Architecture



Does it implement bottom-up attention 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
 - ▶ Fixation traverses from location with highest saliency to lowest
- Hard attention
 - ▶ Binary saliency values
 - ▶ Fixation at the region with saliency = 1
 - ▶ One or very few “salient” objects in a scene
- NN based attention models generally use hard attention



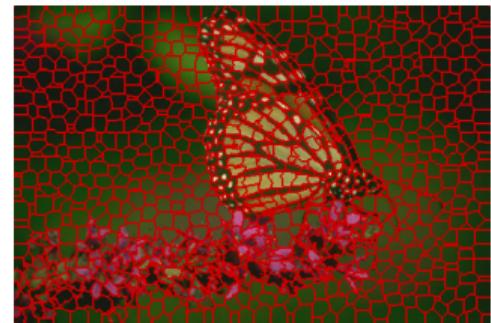
How to train?

Objects and saliency

Saliency-cut algorithm

- Saliency of the nearby pixels should be similar
- 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
 - ▶ Minimize:
 $O(S) = \sum_i (s_i^{new} - s_i)^2 + \sum_{i,j} w_{ij} (s_i^{new} - s_j^{new})^2$
 - ▶ Weights w_{ij} decreases with physical distance
 - ▶ Optimal weights are learned

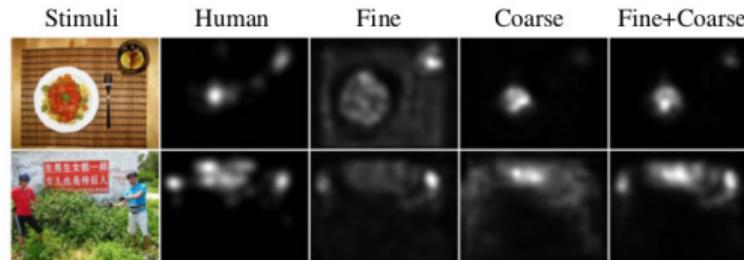
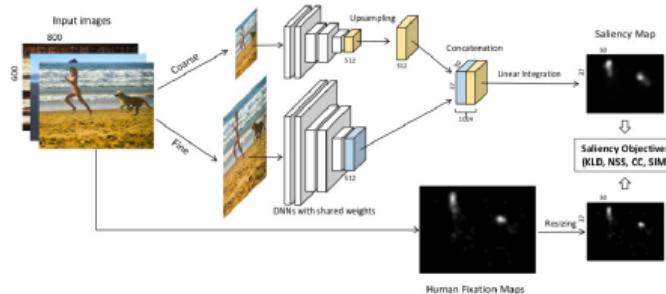
Something like graph-cut algorithm



Graph-cut algorithm (slide deck)

Multi-scale analysis

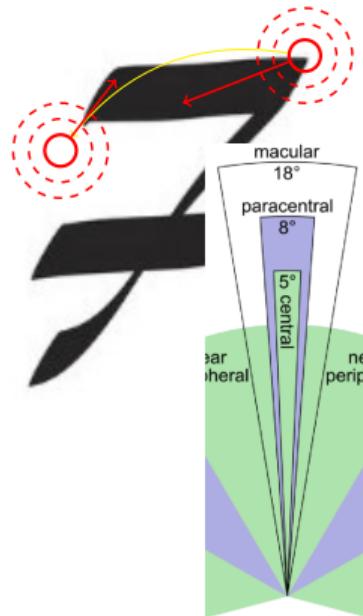
SALICON: Saliency in Context



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

Recurrent Attention Models

Saliency is dynamically constructed

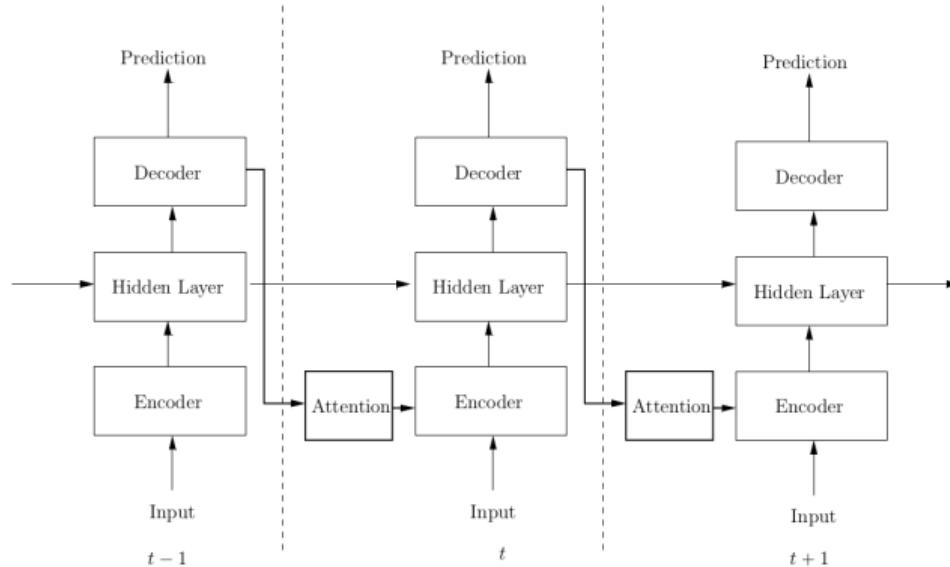


- We look at a small part of a scene at a time
- Where we look at next depends on what we see
 - ▶ Macular/peripheral vision guides the direction of eye movement
 - ▶ ... plus, the task at hand
- Saliency map of a scene is not computed in one go
 - ▶ Constructed dynamically over time
 - ▶ As and when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

Recall EdPuzzle Assignment – Visual Attention: Just in time representation

Attention-based RNN Architecture

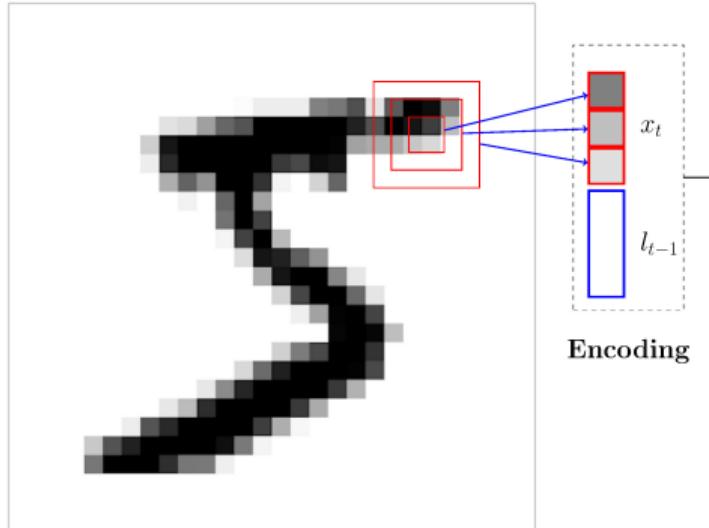
AB-RNN



- RNN and the “Attention” module are trained together

Implementation example

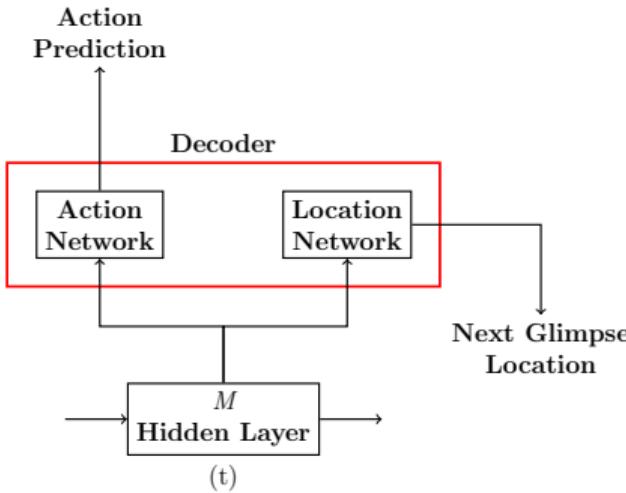
Attention-based object recognition



- 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

Decoder



- Each of Action and Location Networks is an NN
- “Action” can be different in different contexts:
 - ▶ Predicting the object
 - ▶ Locating a target
 - ▶ Navigating (car, pedestrian, drone, ...)
 - ▶ ...

Training

Reinforcement Learning

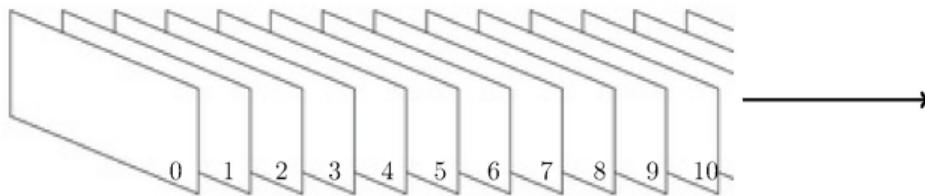
- Model for “hard attention”
- Training for optimal saccades
 - ▶ Training based on back-propagation does not work
 - ▶ Reinforcement learning used
 - ▶ Reward after each time-step
- In the case of object recognition
 - ▶ Reward $r_t = 1$ if the object is classified correctly at time step t
 - ▶ $r_t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

[Reinforcement learning \(tutorial slides\)](#)

Discussions

- Attention and object recognition goes hand-in-hand
 - ▶ Example of task-based attention
 - ▶ 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

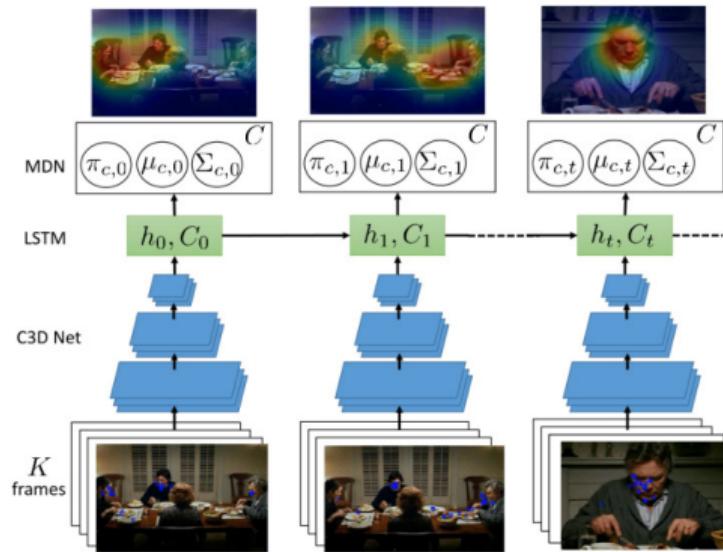


Why processing video frame by frame does not work ?

- 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

- 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 06-02

End of Module 06-02

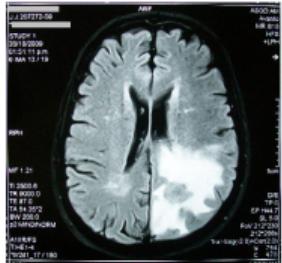
Biological Vision and Applications

Module 07-01: Knowledge Representation

Hiranmay Ghosh

Knowledge required for visual interpretation

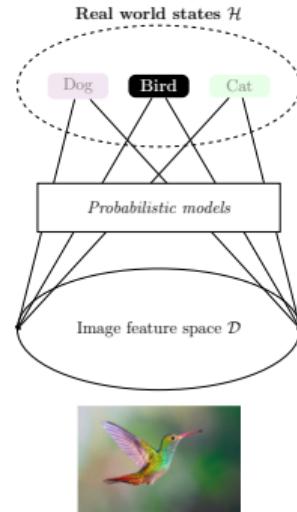
Explicit and implicit knowledge



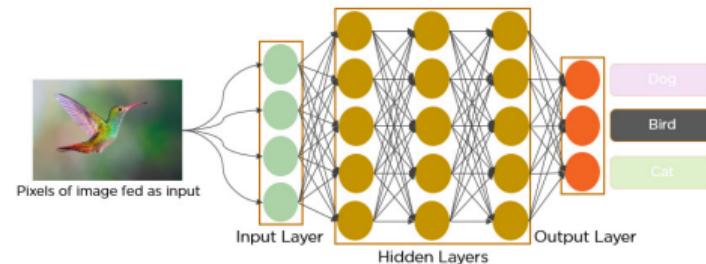
- Domain Knowledge (Ontology)
 - ▶ Example: What makes a lesion different ?
 - ▶ Declarative: can be stated explicitly
 - ▶ Exists independent of processing structure
 - ▶ can be shared
- Knowledge about (image) processing
 - ▶ Example: How to extract the edges from an image?
 - ▶ Procedural: implicit
 - ▶ Encoded as algorithms, neural networks, etc.
 - ▶ Strictly private to the processing structure
- The relation between the two
 - ▶ How to use them together in a problem

Dual representation of knowledge

Same knowledge can be represented either in declarative or in procedural way

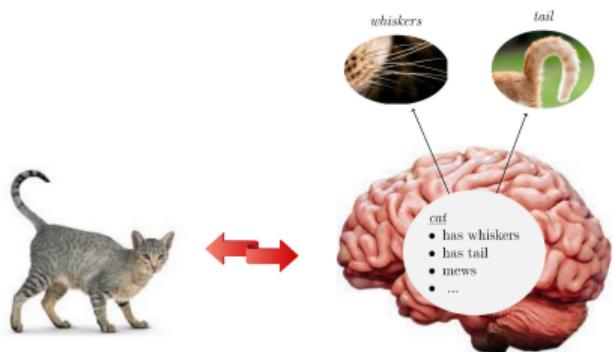


Declarative
Model based approach



Procedural
Data driven approach

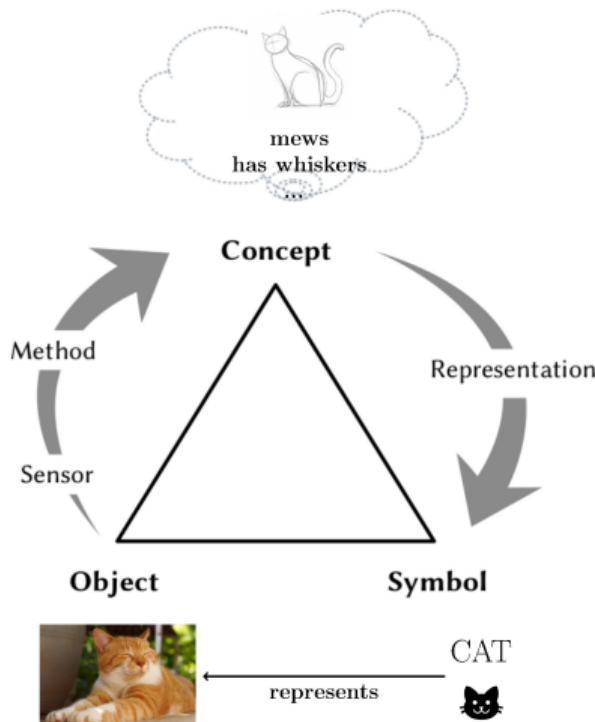
Representational Theory of Mind (RTM)



- A **concept** is a mental model of something
 - ▶ A real-world thing
 - ▶ An internal mental state of the agent
- A name is associated with a concept
 - ▶ For reference while manipulation
- Knowledge is
 - ▶ A collection of named concepts
 - ▶ A set of sentences (propositions) that relate the concepts
 - ▶ Named concepts: cat, tail
 - ▶ Proposition: A cat has a tail

Symbolic representation

The semiotics triangle

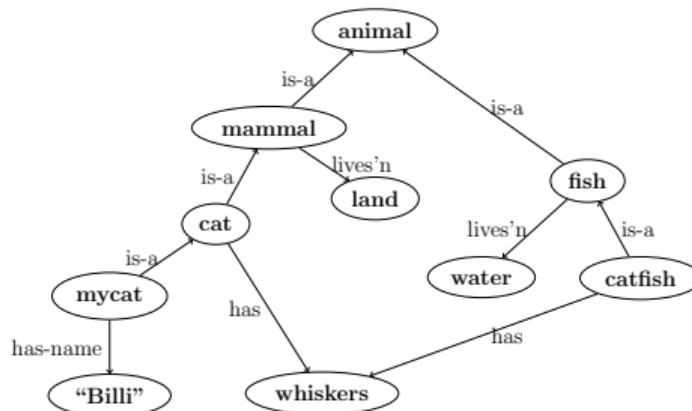


- **Objects** (things): That exist
- **Concepts**: Mental representations (models)
- **Representation**: Symbol to represent a concept (text, icon, speech)

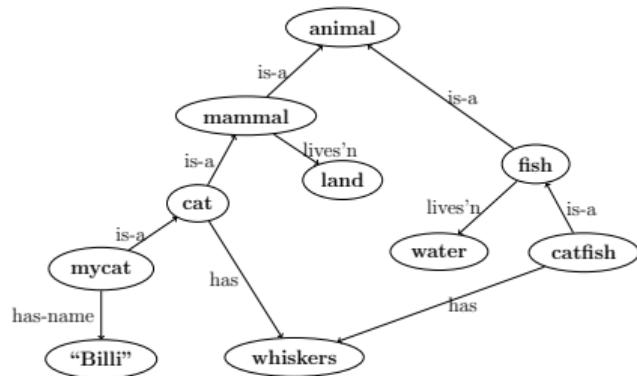
Semantic Network

Compact knowledge representation

- 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 “Billi”
- Equivalently, knowledge is a graph (semantic network)



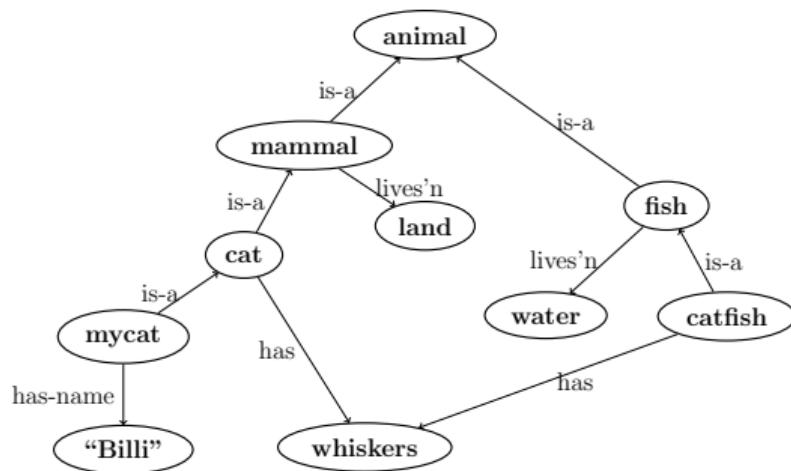
Semantics of “Semantic Network”



- A set of propositions
 - ▶ Each describes a property of a concept
- Structure: *< subject, predicate, object >*
 - ▶ Examples:
 - ▶ cat has whiskers
 - ▶ mycat has-name "Billi"
- A concept can be a class, or an instance
 - ▶ Example: cat, mycat
- A value can be a concept, or a literal
 - ▶ Example: whiskers, "Billi"

Reasoning with Semantic Network

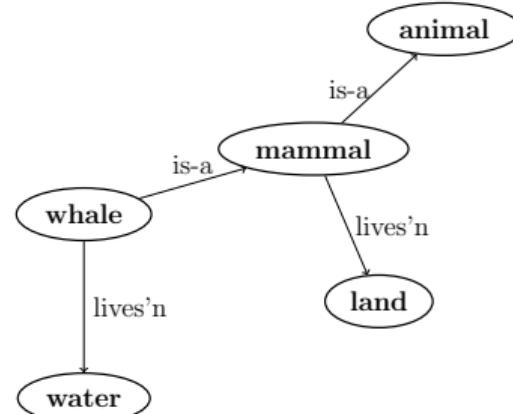
Axioms lead to semantics



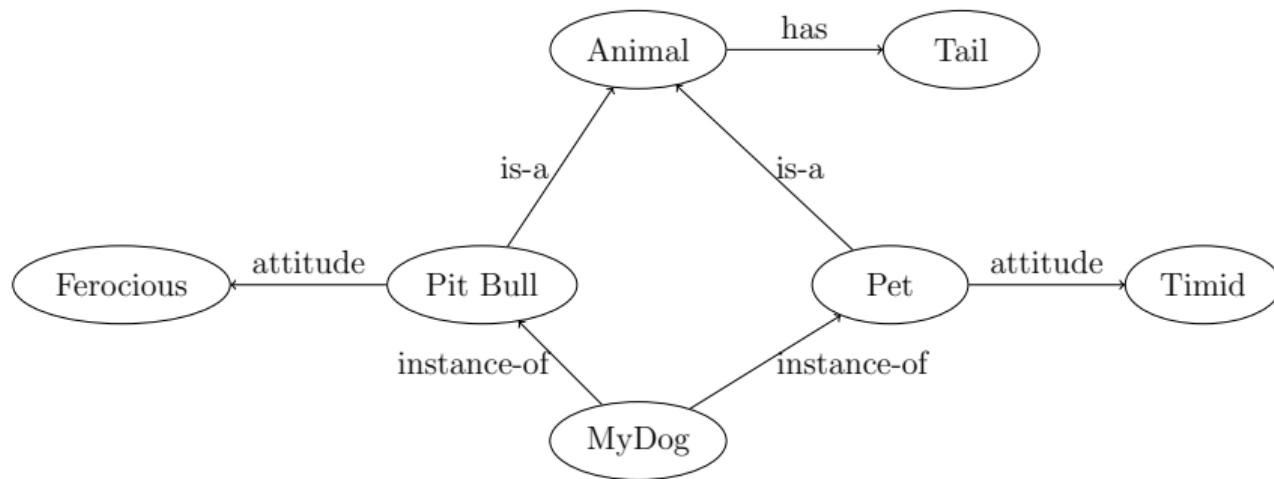
- Requires underlying axioms, e.g.
 - ▶ Property inheritance
 - ▶ If a “is-a” b , then a inherits properties of b
 - ▶ “is-a” is transitive
 - ▶ If a is-a b , and b is-a c , then a is a c too
 - ▶ These axioms make a semantic network efficient (compact)
 - ▶ There can be other domain-specific rules

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



MyDog has a tail

What's about MyDog's attitude ?

Properties of Semantic Network

- A semantic network is extremely flexible
 - ▶ An informal description of a domain (in its 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

Resource Description Framework (RDF)

W3C Recommendation – for representing interconnected data on the web

- Data (resources) can be distributed over the web
 - ▶ Knowledge is an interconnection (relations) of this distributed data
- Each resource is identified with a IRI
- Follows semantic network model
 - ▶ An RDF graph is a set of RDF sentences ⟨ subject, predicate, object ⟩
- A predicate is also a “resource”
 - ▶ 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”
 - ▶ I said “cat is an animal”
 - ▶ S1: ⟨ cat, is-a, animal ⟩. S2: ⟨ I, said, S1 ⟩
- Some axioms are inbuilt – additional semantics can be defined with 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



Quiz 07-01

End of Module 07-01

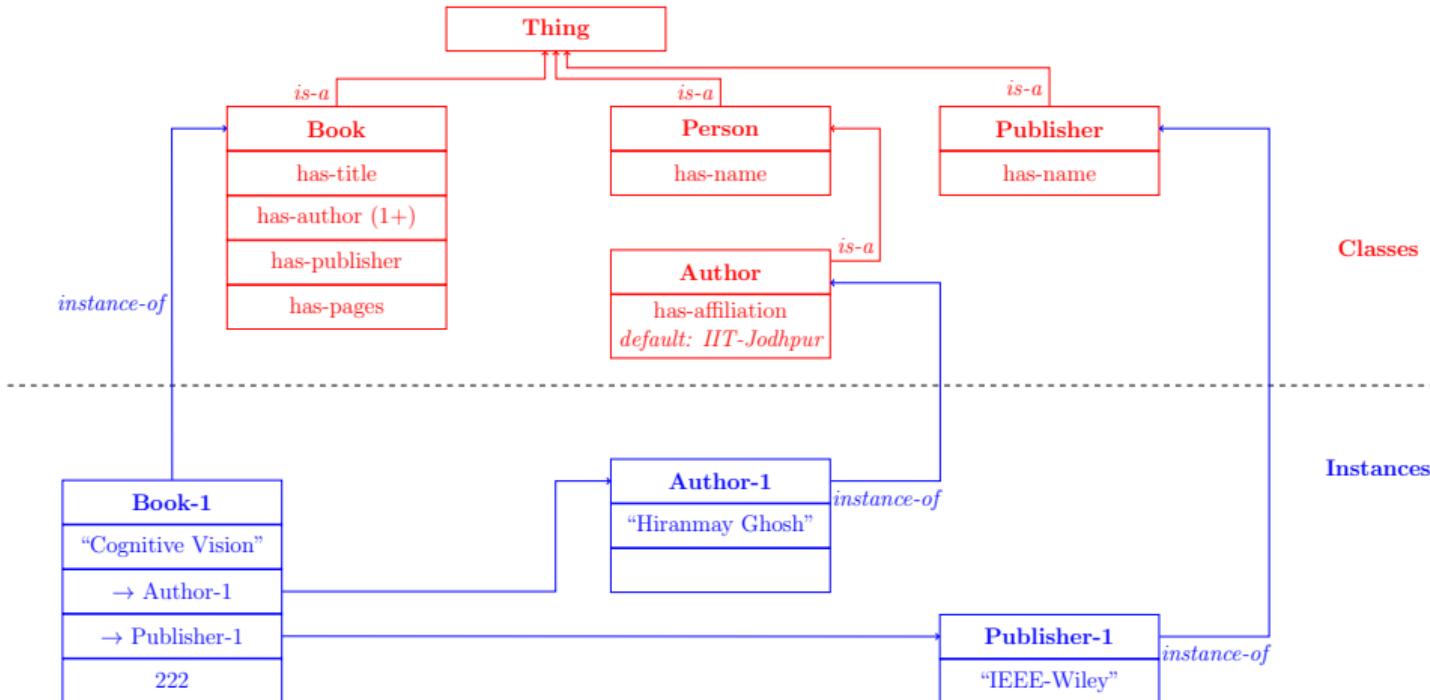
Biological Vision and Applications

Module 07-02: Frame-based Knowledge Representation

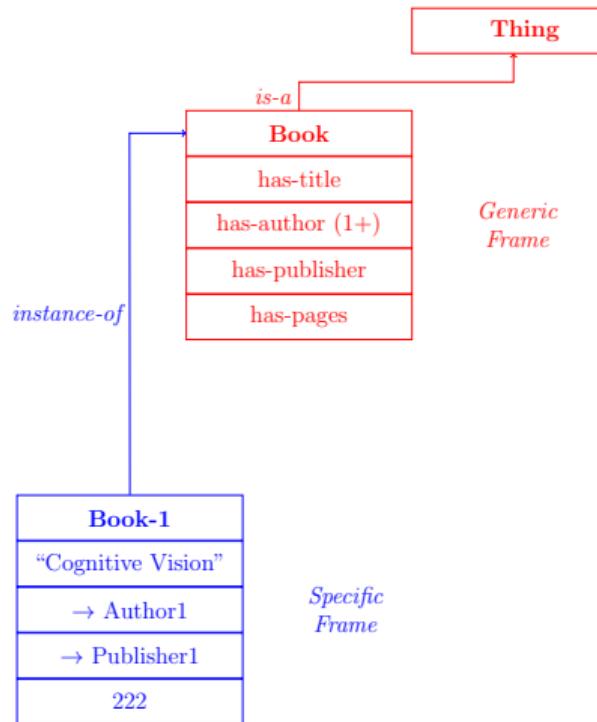


Hiranmay Ghosh

Frame-based representation



Frames, slots and fillers



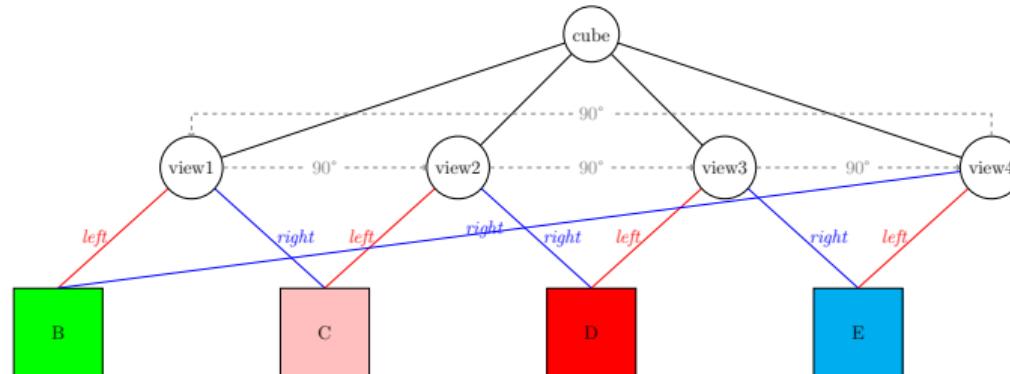
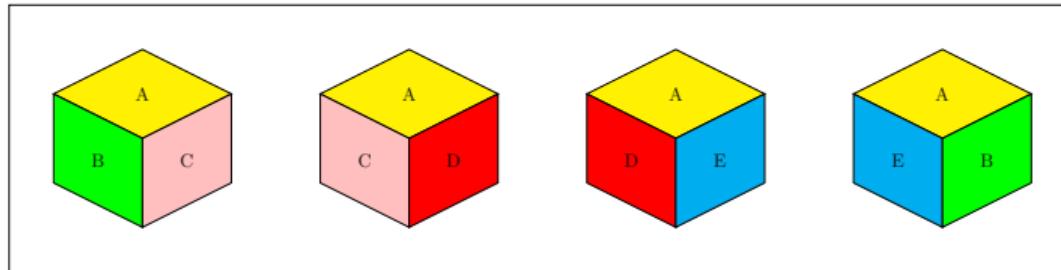
- A frame has a label
- A frame consists of one or more **slots** (attributes)
- A slot contains a **filler** (value)
 - ▶ Reference to another frame
 - ▶ Literal
 - ▶ not specified
- A frame inherits attributes and default values of its parent
- Value restrictions
 - ▶ Data types / range
 - ▶ Cardinality

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

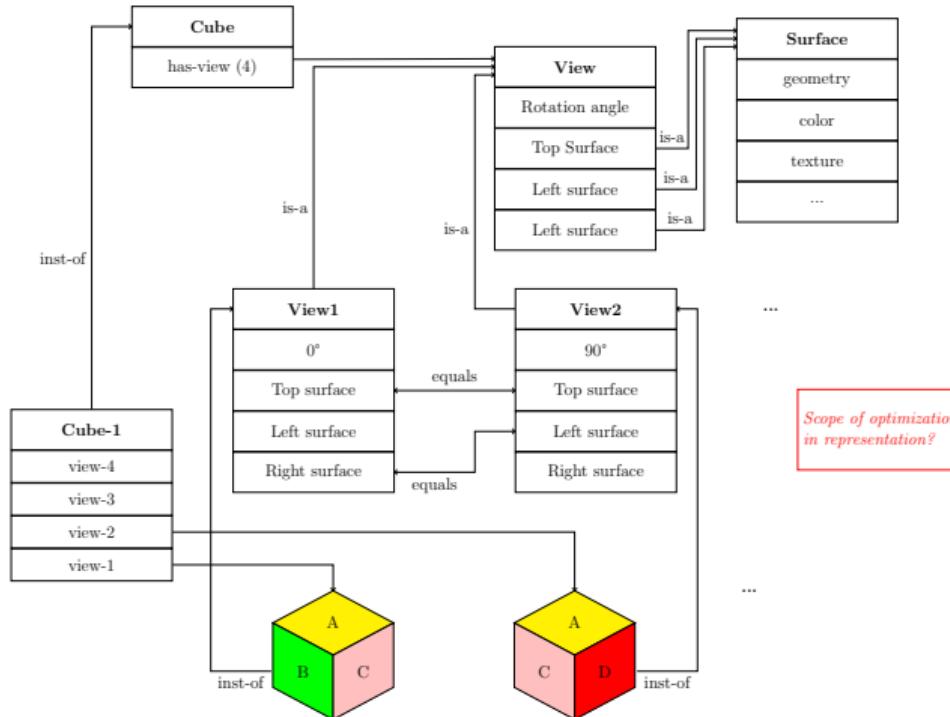
Frame based representation & Visual cognition

Visual events and viewpoints

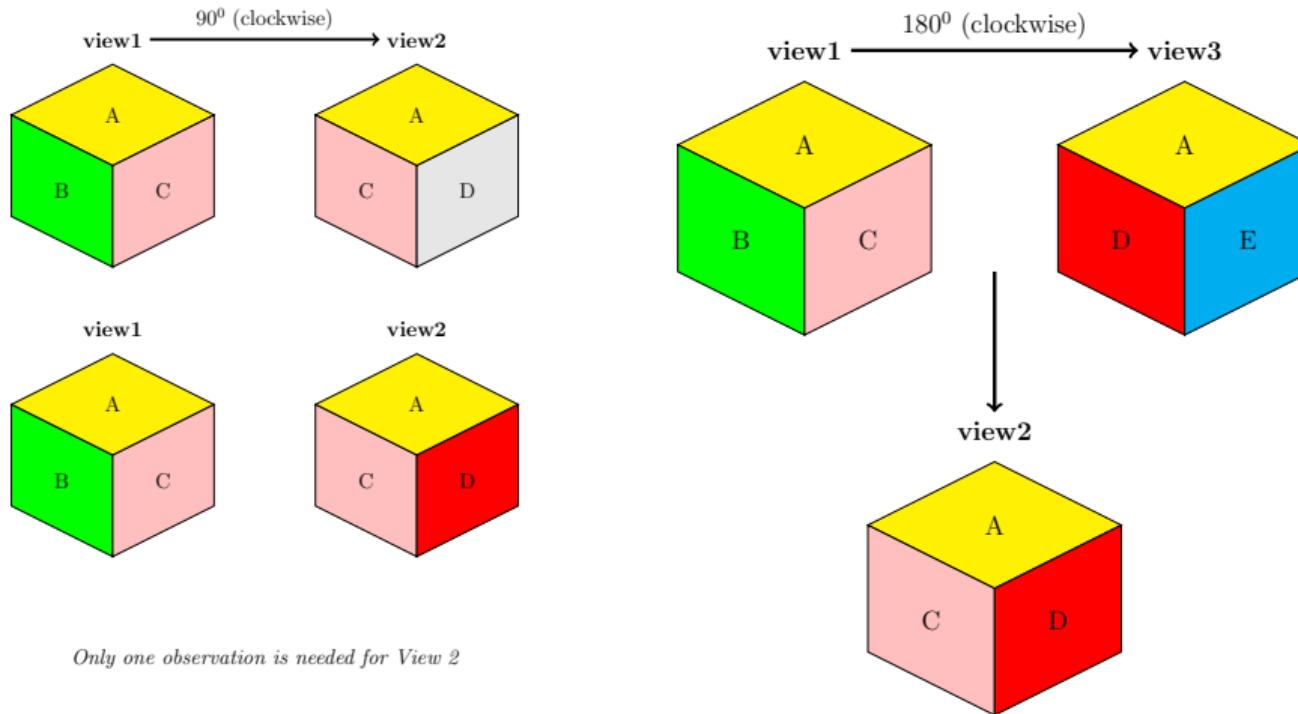


Frame based representation & Visual cognition

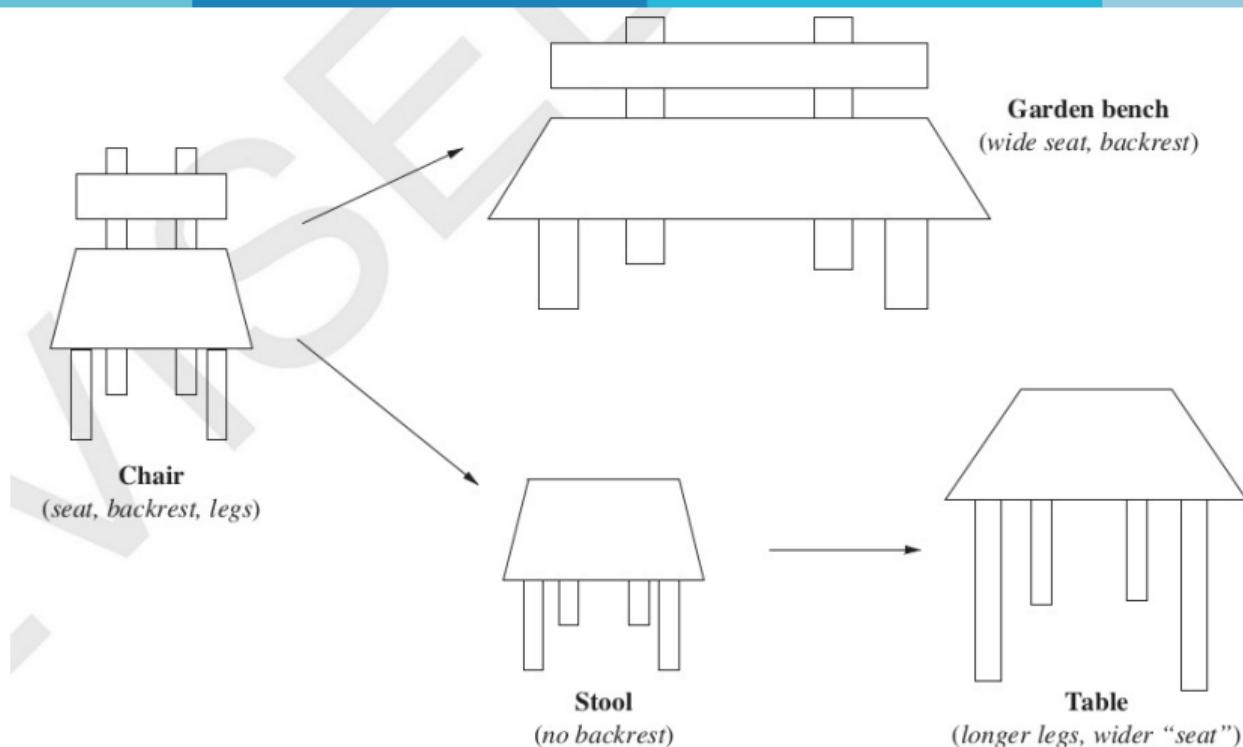
Viewpoints: frame-based representation



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

Biological Vision and Applications

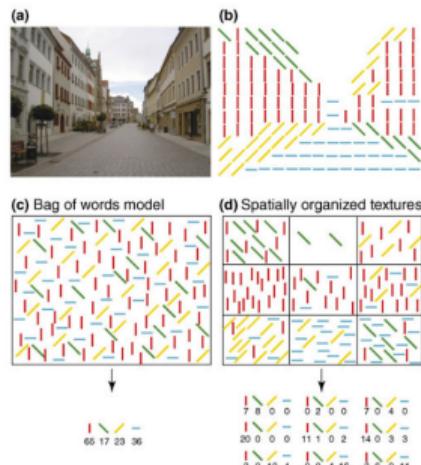
Module 07-03: Part-based recognition

Hiranmay Ghosh

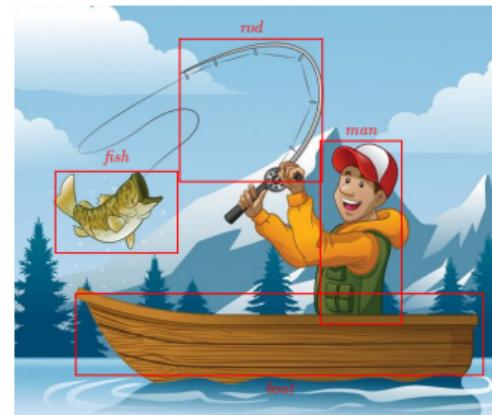
Visual recognition

Holistic vs. Part-based

Holistic representation



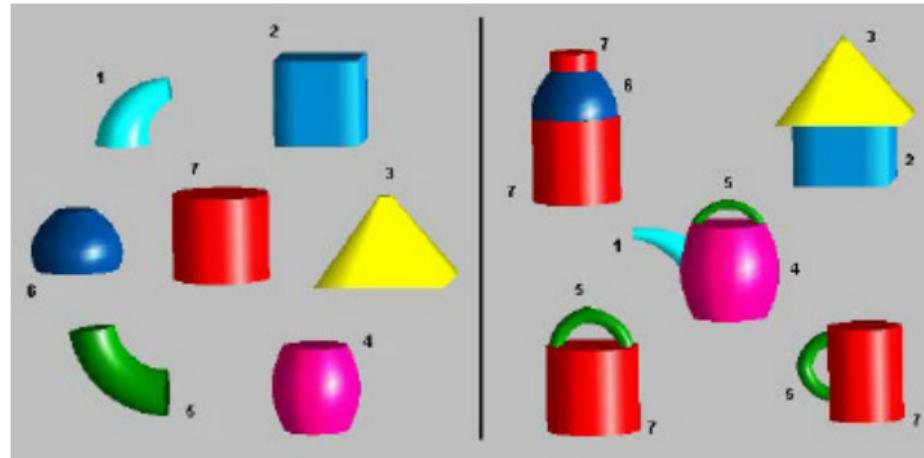
Structured (part-based) representation



EdPuzzle: Dual Process Theory

Part-based object recognition

An object is composed of some elementary 3D parts

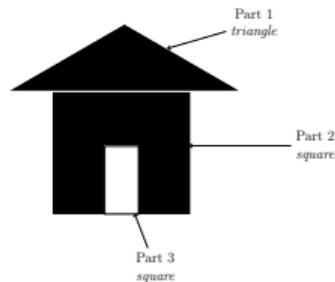


geons

- Composability as in natural language
 - ▶ Parts = visual words (elementary units for visual recognition)

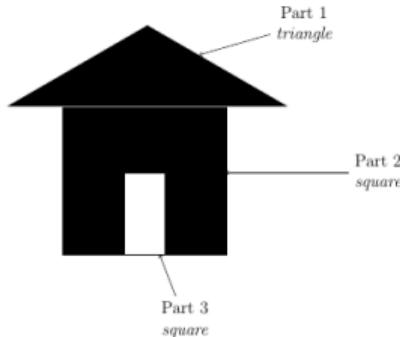
How to define a part?

- Operational definition: Parts are what can be detected by part detectors
- Definition based on principle of simplicity:
Parts are polynomial surfaces approximating closed, non-overlapping image regions that optimally partition the image (MDL)



The elementary parts (geons) are characterized by their shapes, size, colors and textures.

Perceptual organization of objects



Each part (geon) is characterized by shape and relative size

Structural relations:

Part-1 *above* Part-2

Part-3 *contained-in* Part-2

- Geon-diagrams with similar structural composition represent same kind of objects
 - ▶ Appearance model for a geon: (shape, rel size)
 - ▶ Relations between geons (structure)
- Geon-diagrams with identical geons, but with different structural composition may represent different kind of objects

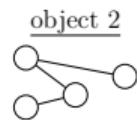
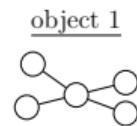
• Semantics:

▶ A *above* B → B *supports* A

▶ A *contained-in* B → A *part-of* B

Object recognition with perceptual model

An object can be represented like a graph (geon diagram)



...

Each node represents a geon

- Characterized by shape and rel size

Each edge represents a relation between two geons

- Above, contained-in, etc.



Observed object Graph

Which object graph explains the observation best ?

Graph matching: ... (short paper) *

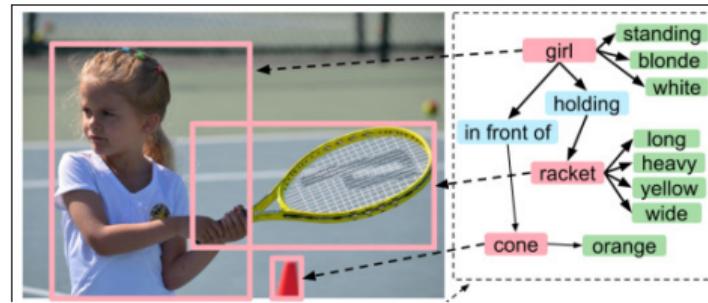
Probabilistic model and learning

- Let G be a finite vocabulary of geons, R a finite set of relations
- Each object is characterized by a probabilistic geon-graph: (V, E)
 - ▶ $v \in G$: a geon, characterized by a probabilistic appearance model (*type, size*)
 - ▶ $e \in R$: Probabilistic relations between a pair of vertices
 - ▶ Parameters can be learned over a large number of observations
- An observation is also a geon-graph
- Inference: Which object explains the observed graph the best ?

- Probabilistic model & learning: Crandall & Huttenlocher. Weakly Supervised Learning of Part-Based Spatial Models ...
 - Application to neural network: Krause, et al. Learning Features and Parts for Fine-Grained Recognition ...

Activity recognition and Scene Graph

- Girl playing tennis
- objects + locations + interactions

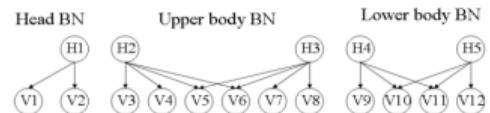
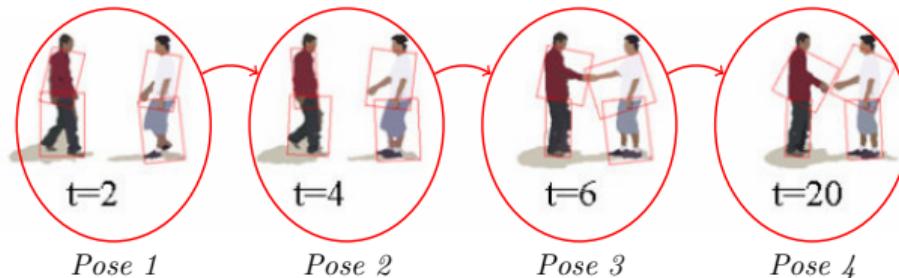


Chang, et al. Survey of scene graphs

Johnson, et al. Image Retrieval using scene graph

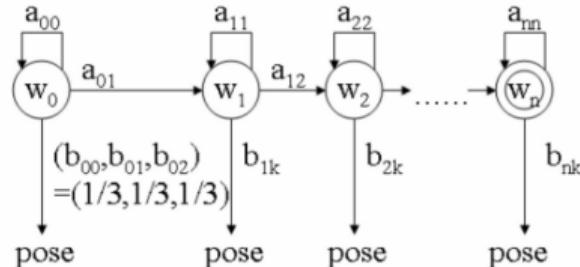
Event recognition

Extending Scene Graph in time



Definitions for nodes:

- H1: Head = {0:front, 1:left, 2:right, 3:rear}
- H2: ArmV = {0:high, 1:mid-high, 2:mid-low, 3:low}
- H3: ArmH = {0:withdrawn, 1:intermediate, 2:stretched}
- H4: LegV = {0:high, 1:middle, 2:low}
- H5: LegH = {0:withdrawn, 1:intermediate, 2:stretched}
- V1: angle of vector from head to face
- V2: ratio of face to head
- V3: y-position of hand blob
- V4: y-pos of far most point of upper body
- V5: upper body ellipse ratio
- V6: upper body ellipse rotation
- V7: x-position of hand blob
- V8: x-pos of far most point of upper body
- V9: y-pos of far most point of lower body
- V10: lower body ellipse ratio
- V11: lower body ellipse rotation
- V12: x-pos of far most point of lower body



Hidden Markov Model (HMM) representation of an event

Quiz



Quiz 07-03

End of Module 07-03

Biological Vision and Applications

Module 07-04: Knowledge representation for visual cognition



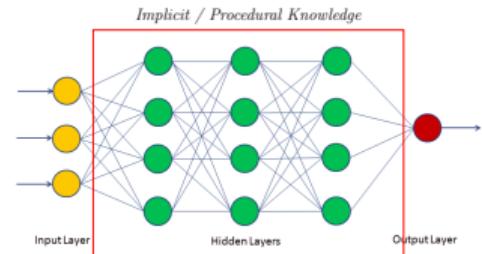
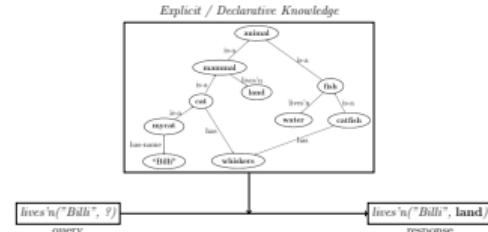
Hiranmay Ghosh

Dual Process Theory

Two processing paradigms

- Knowledge-driven approach (top-down)
 - ▶ Motivated by AI research
 - ▶ Examples: Logic systems, frame systems, ...
 - ▶ Symbolic Systems

- Data-driven approach (bottom-up)
 - ▶ Motivated by neuro-sciences / ML research
 - ▶ Examples: Learnt classifiers, Neural networks
 - ▶ Emergent Systems



Symbolic approach is traditionally known as the "cognitive approach"

Comparison of symbolic and emergent system approaches

Symbolic Systems	Emergent Systems
Formal representation: sharable	Informal representation: private
- Human understandable / creatable	- Not human understandable
Structured: can combinatorially generalize	Monolithic: cannot generalize
Inflexible: cannot discover new concepts	Flexible: can find new patterns in data
Deliberative reasoning: formal methods, slow	Intuitive Understand: informal methods, fast
Brittle: less tolerant to noisy data	Robust: more tolerant to noisy data
Explainable	Not explainable
Model-based	Model-less

See table 6.1 in textbook

How does the human mind work?

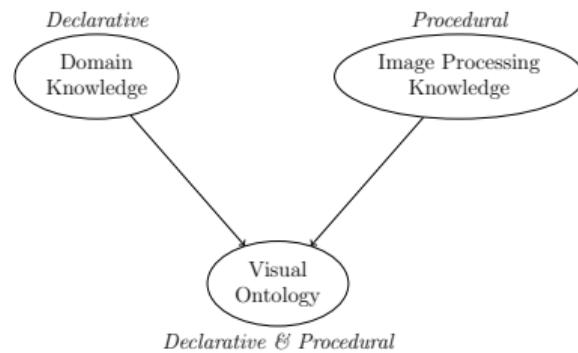
Dual process theory

- Symbolic approach: formal, knowledge driven, explainable, accurate, slow
 - ▶ More suitable for cognitive tasks
 - ▶ I need to go to the institute – how do I go?
- Emergent system approach: informal, data-driven, not explainable, inaccurate, fast
 - ▶ More suitable for perceptual processing
 - ▶ Is it an apple or a banana?
- Dual process theory: Human mind uses both the approaches
 - ▶ Parallel model: fast and slow thinking occur simultaneously
 - ▶ ... and may conflict
 - ▶ Default-Interventionist model: fast thinking generates intuitive responses
 - ▶ Subsequent slow thinking process may or may not deliberate on them

[Dual process theory \(short article\)](#)

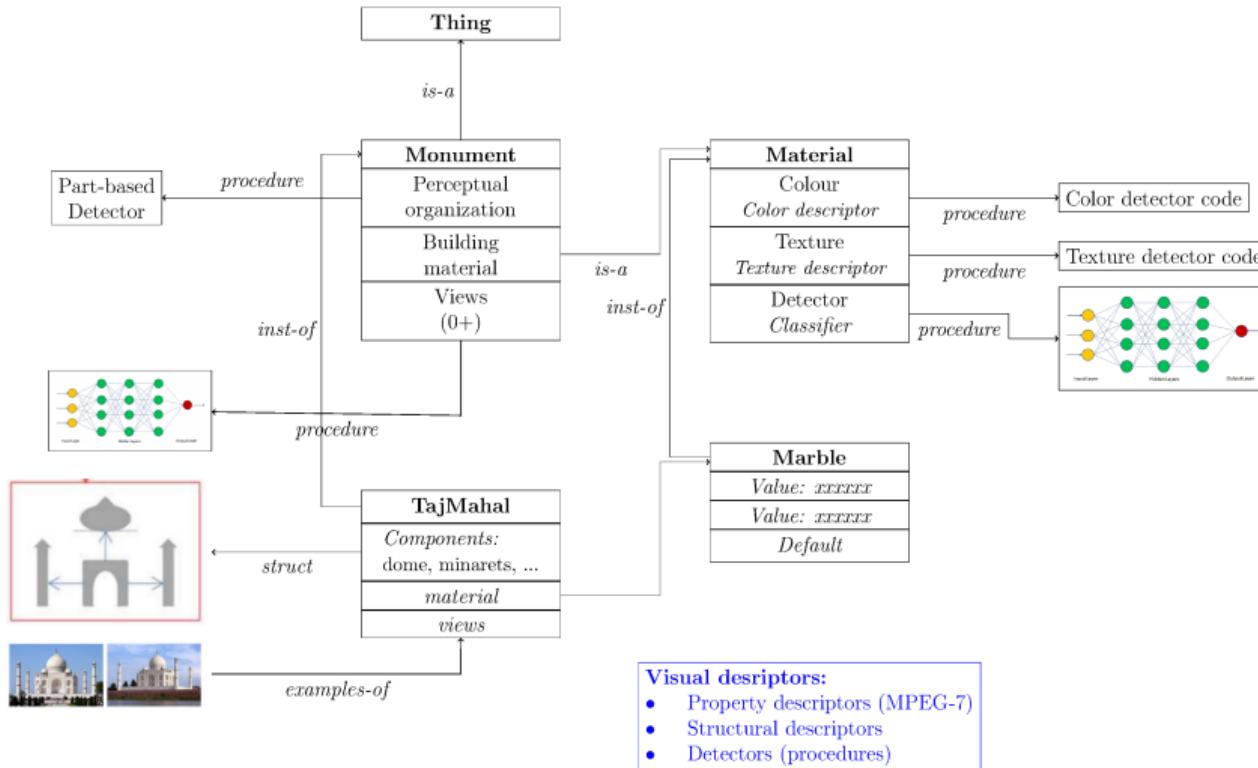
Framework of visual knowledge presentation

- **Ontology** =
 - ▶ A set of concepts C
 - ▶ A set of relations R
 - ▶ A set of axioms, e.g. transitivity, reflexivity, etc.
 - ▶ Two partial orders \succ_C and \succ_R define concept and relation hierarchies



Example of visual property specifications in a frame-based system

Multiple modes of specification in a unified framework



More of structural description

Point events and their relations



Temporal:
 $\overline{R \text{ before } B}$

Spatial:
 $R \text{ north / above } B$
 $R \text{ west / left } B$

Inverse relations:
 $R \text{ before } B \equiv$
 $B \text{ after } R$

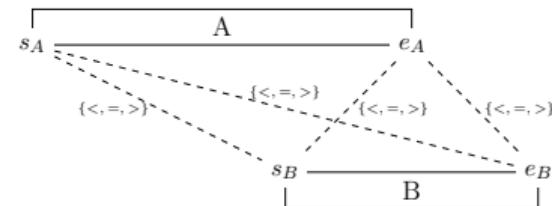
$R \text{ north } B \equiv$
 $B \text{ south } R$

- The relations are normative
 - ▶ based on convention ... lacks semantics
- Events are seldom point events
 - ▶ finite spatial and temporal extension

Temporal relations between finite events

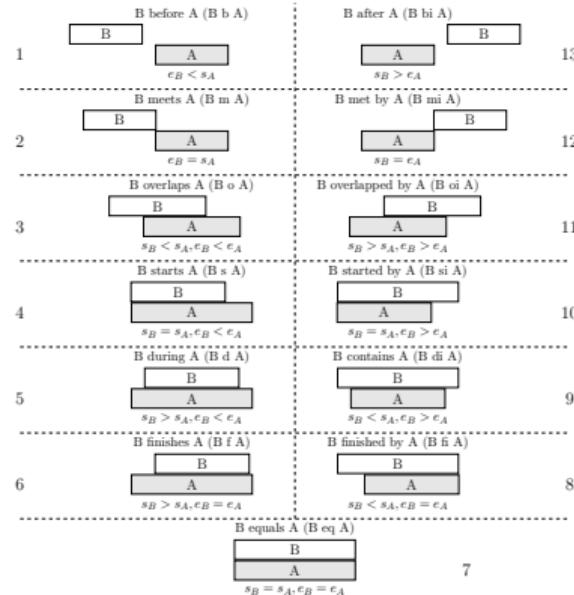
Allen's temporal relations

- 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 have 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

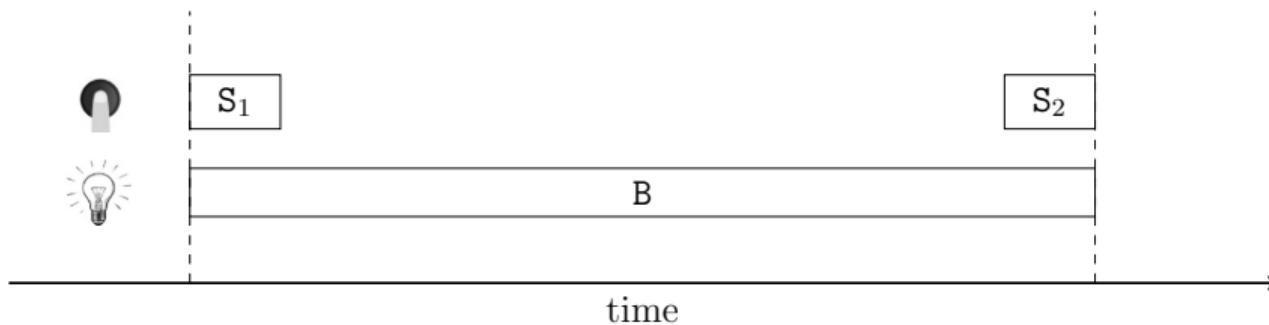


1. $e_B < s_A$: B *before* A
2. $e_B = s_A$: B *meets* A
3. $s_B < s_A, e_B < e_A$: B *overlaps* A
4. $s_B = s_A, e_B < e_A$: B *starts* A
5. $s_B > s_A, e_B < e_A$: B *during* A
6. $s_B > s_A, e_B = e_A$: B *finishes* A
7. $s_B = s_A, e_B = e_A$: B *equals* A

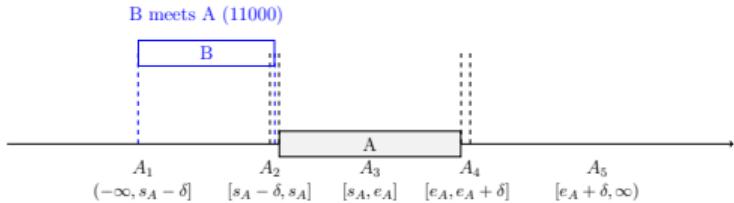
Allen's temporal relations

Example

- (a) S_1 before S_2
- (b) S_1 starts B
- (c) S_2 finishes B



Binary encoding



B before A:	10000	B after A:	00001
B meets A:	11000	B met by A:	00011
B overlaps A:	11100	B overlapped by A:	00111
B starts A:	01100	B started by A:	01111
B during A:	00100	B contains A:	11111
B finishes A:	00110	B finished by A:	11110
B equals A:	01110		

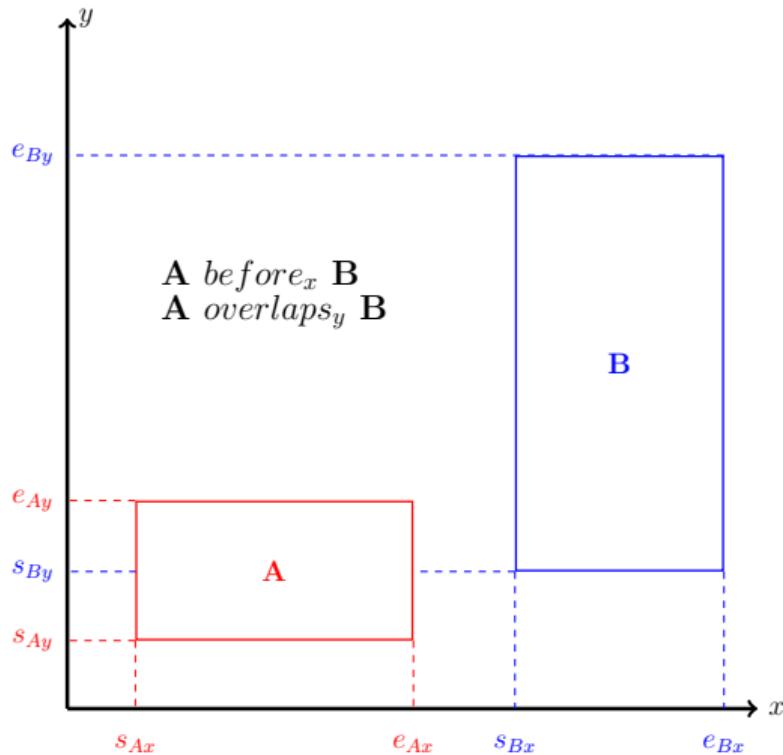
- There can be small error in determining the exact boundaries of an event
 - ▶ “B meets A” may be confused with “B before A” or “B overlaps A”
- Hamming distance between the binary strings signify closeness of the relations

Papadias. Approximate Spatio-temporal Retrieval

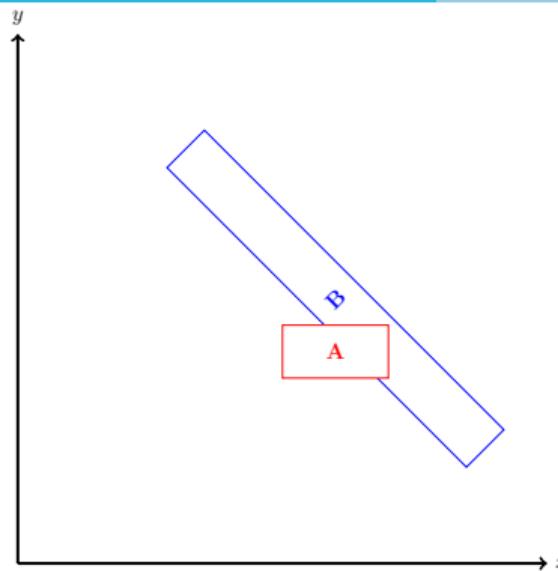
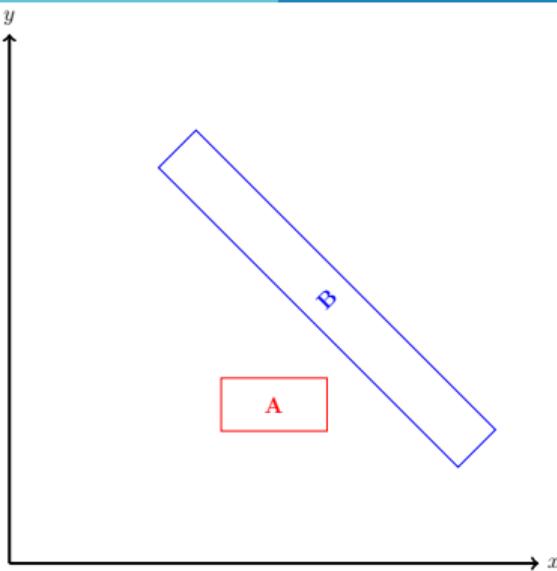
Allen's relations

Extension to spatial dimensions

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



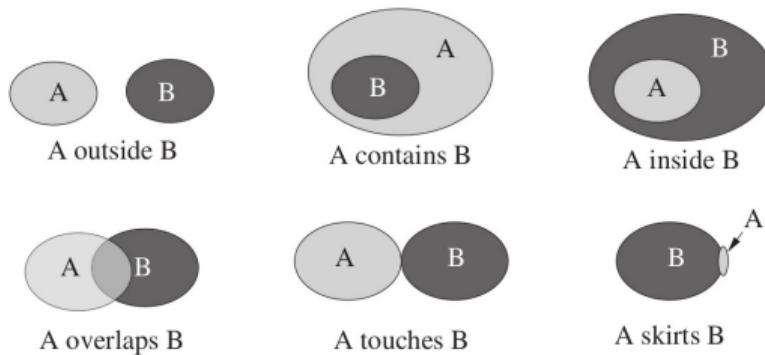
Ambiguity: Allen's relations extended to multi-dimensional space



- 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



Quiz 07-04

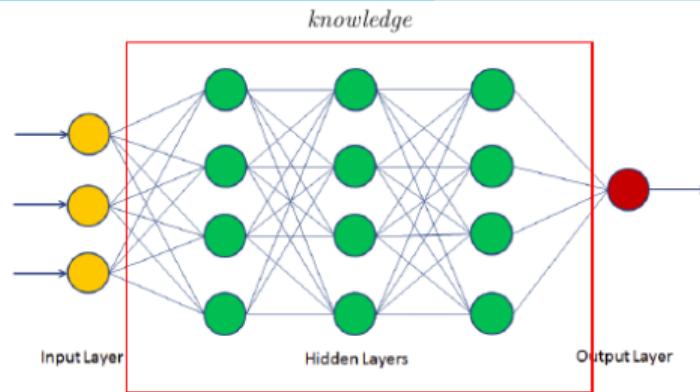
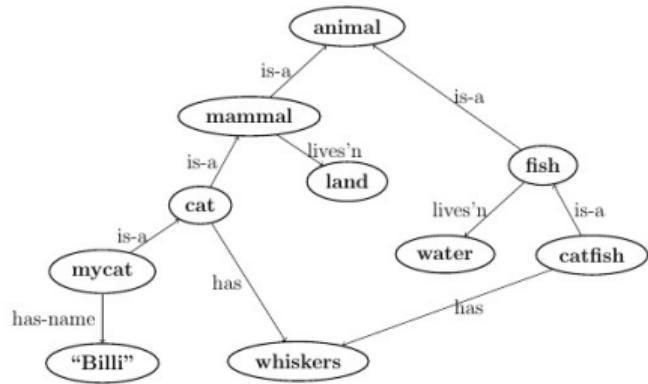
End of Module 07-04

Biological Vision and Applications

Module 07-05: Graph Neural Networks

Hiranmay Ghosh

Explicit knowledge vs. Implicit knowledge



- **Explicit knowledge**

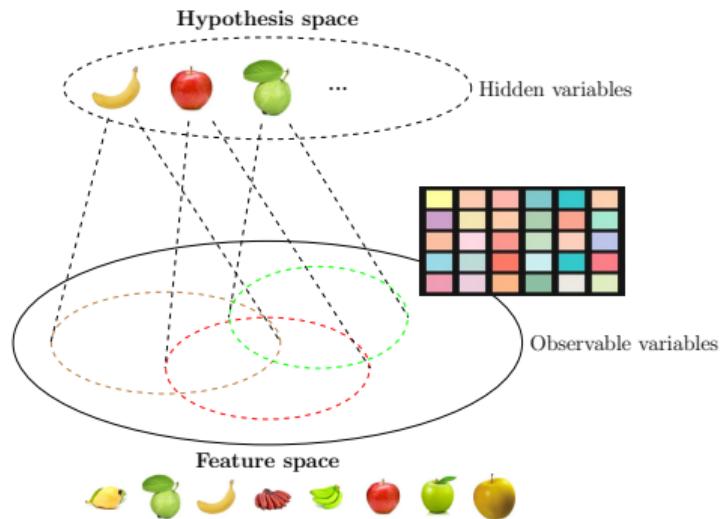
- ▶ Model based
- ▶ Inductive generalization
- ▶ Slow
- ▶ Good for reasoning

- **Implicit knowledge**

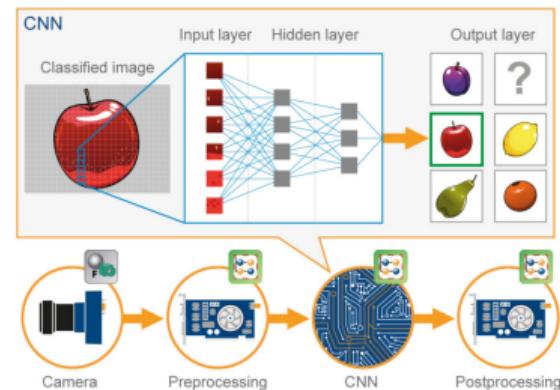
- ▶ Emergent knowledge
- ▶ No generalization
- ▶ Fast
- ▶ Good for understanding

Model-based Reasoning vs. Model-less understanding

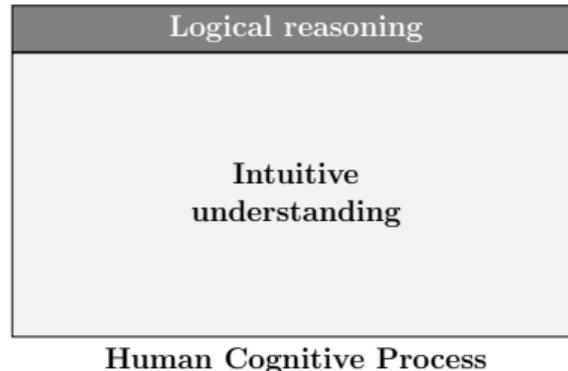
Model-based reasoning



Model-less understanding



Dual Process Theory



- Interaction?
- Reasoning = asking questions

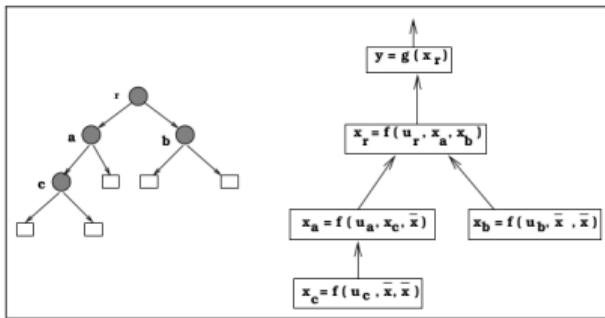
- Can we have reasoning coupled with understanding
 - ▶ “Intersection of deep learning and structured approaches”
 - ▶ Inductive generalization is the key to AI
 - ▶ Structure with emergent knowledge

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

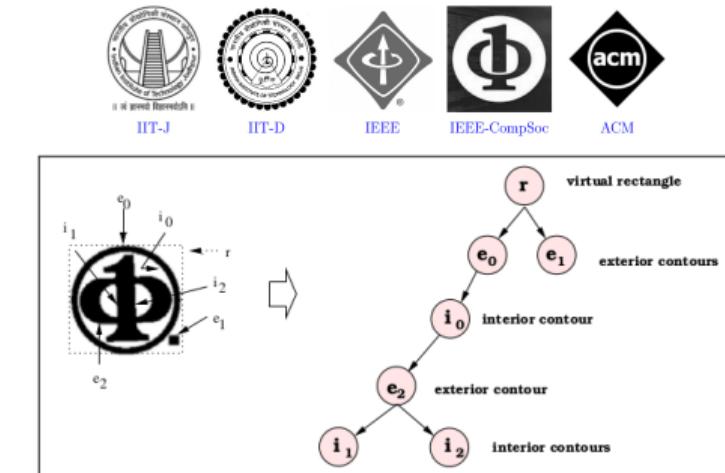
[Comments on the position paper](#)

Combining structural analysis and machine learning

Part-based recognition of logos (1998)



- Functions $f()$, $g()$ realized as NN
 - ▶ Trained with large data set
- Identical property descriptors u_x

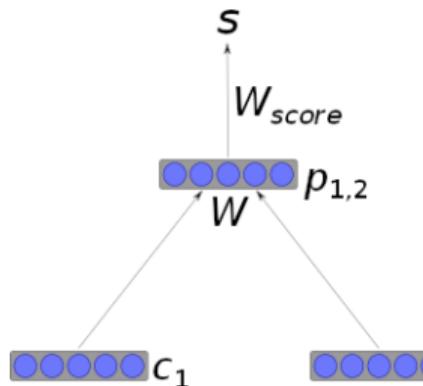


- Fairs better than a single NN

Frasconi, et al. A General Framework for Adaptive Processing of Data Structures

Frasconi, et al. Logo Recognition by Recursive Neural Networks

Recursive Neural Network (RvNN)



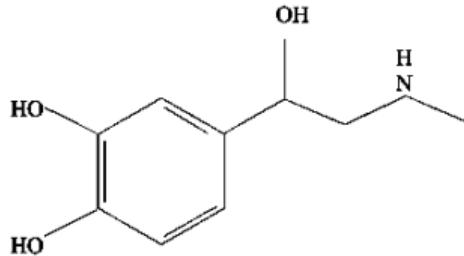
- c_1, c_2 are inputs
 - ▶ Vectors of dimension n
- $p_{1,2}$ is the output
 - ▶ Also a vectors of dimension n
- $p_{1,2} = \tanh(W[c_1; c_2])$, where
 - ▶ W is a learned $n \times 2n$ weight matrix

- Limitations
 - ▶ Data must be organized as a binary tree
 - ▶ Data flow is one way – leaf to root

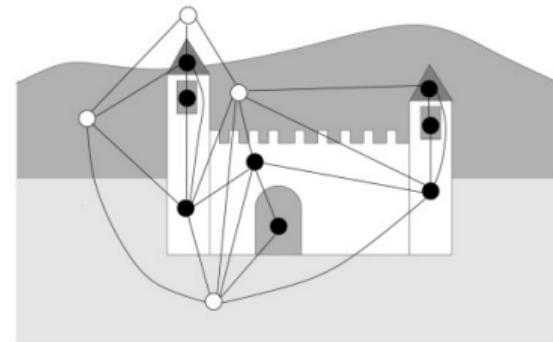
Graph Neural Network

Data is represented as a graph

- Graph focussed Application
 - ▶ What is the property of the molecule ?
 - ▶ $output = \tau(G)$

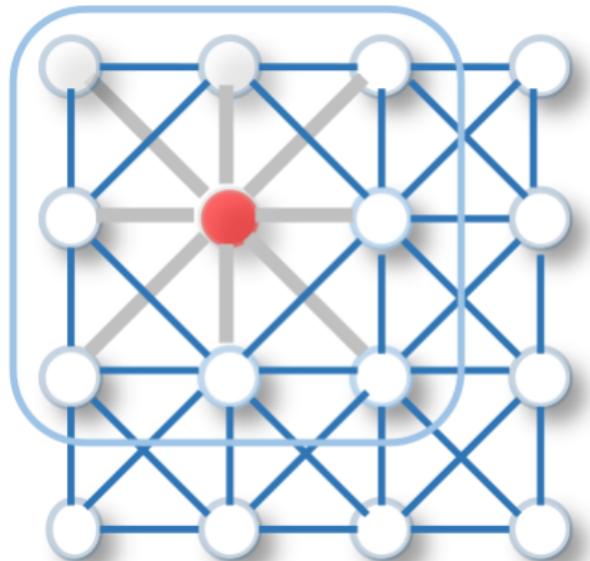


- Node focussed Application
 - ▶ What does each of the nodes represent ?
 - ▶ $output = \tau(G, n)$

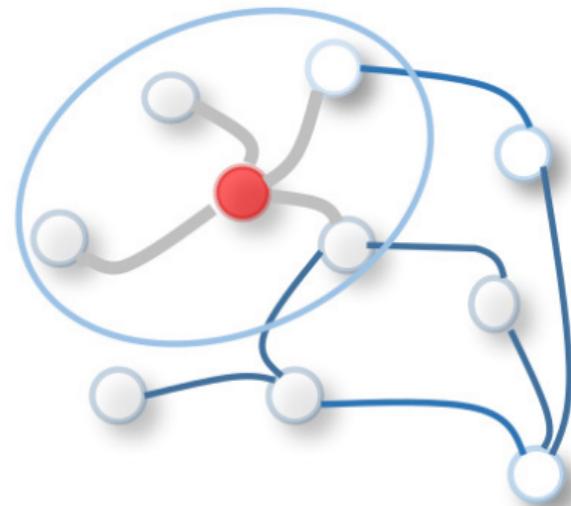


Scarselli, et al. Graph Neural Network (2009) *

2D Convolution vs. Graph Convolution



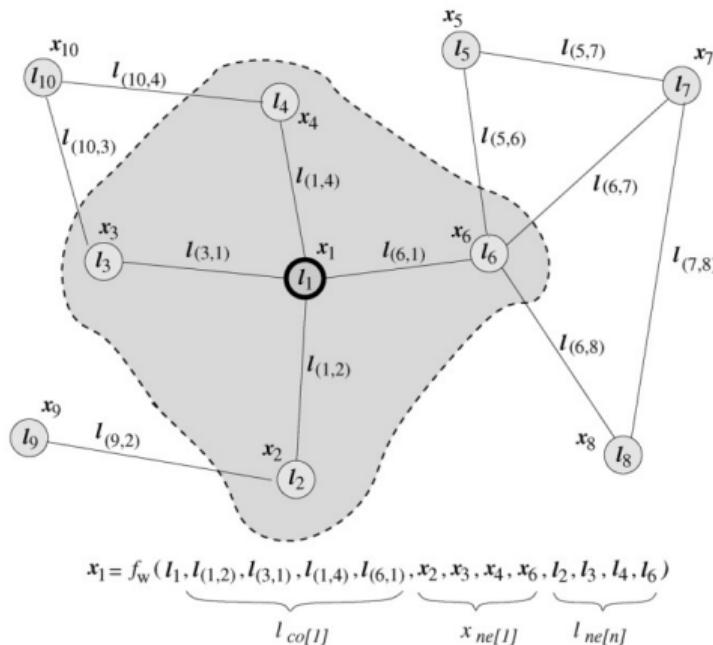
2D Convolution



Graph Convolution

Convolutional Graph Neural Network

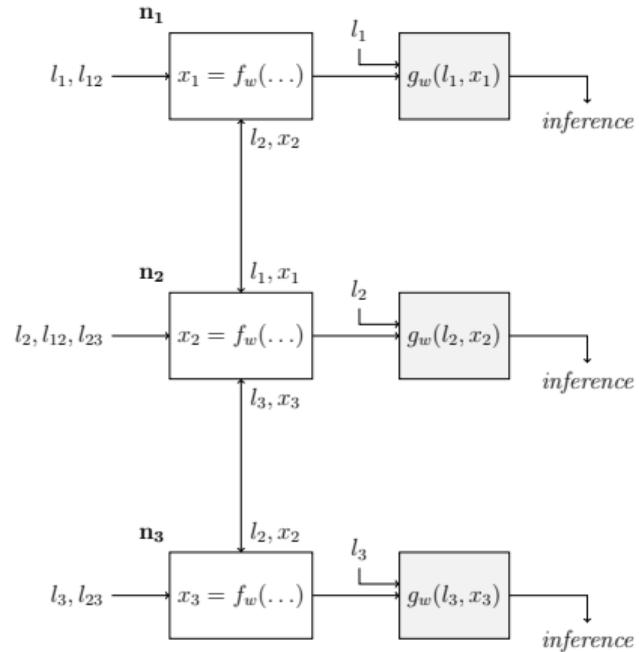
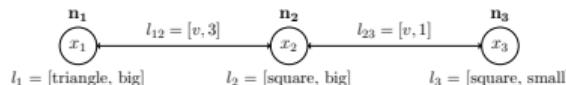
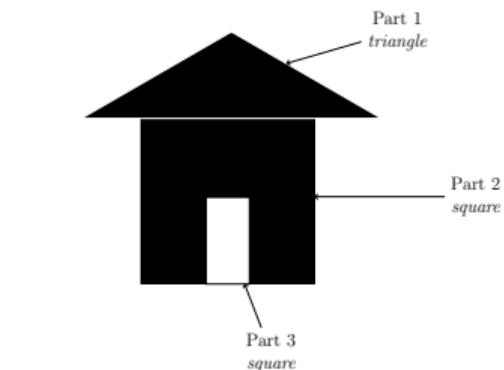
Processing model



- Nodes have identical property (feature) descriptors (I_n)
 - ▶ e.g. color, texture, shape
- Edges have identical property descriptors (I_{mn})
 - ▶ e.g. distance between the center of gravities of the nodes
- State of a node: x_n

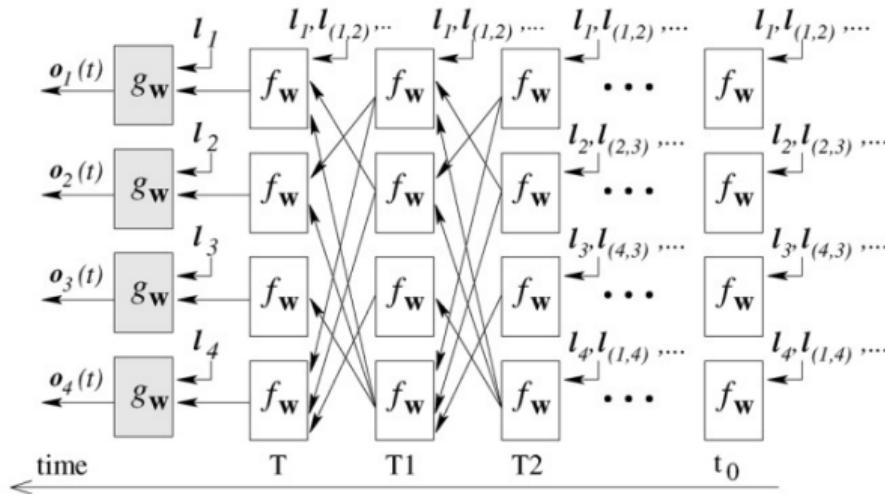
Graph Neural Network

Processing model: Example



Graph Neural Network

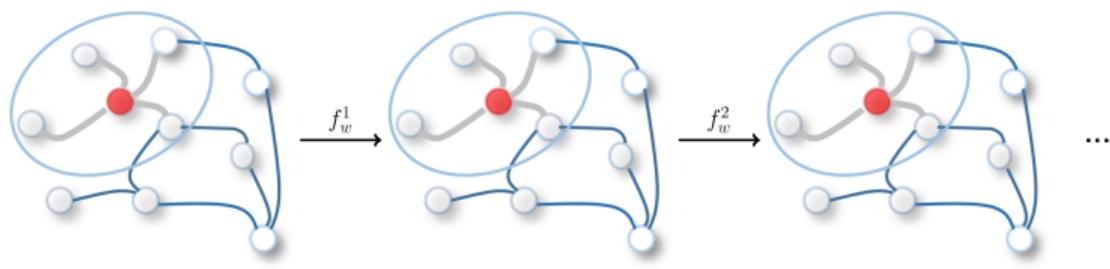
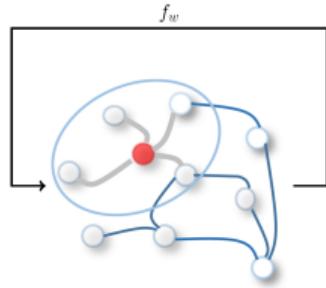
Recurrent Processing



- $f_w()$ and $g_w()$ are learned with training
- Take the output after several recursions
 - ▶ Is the system guaranteed to go into a steady state after a finite number of iterations?

Recursive & Convolutional Graph Neural Network

Rec-GNN & Conv-GNN



Recursive GNN:

Same weights
at every time-stamp

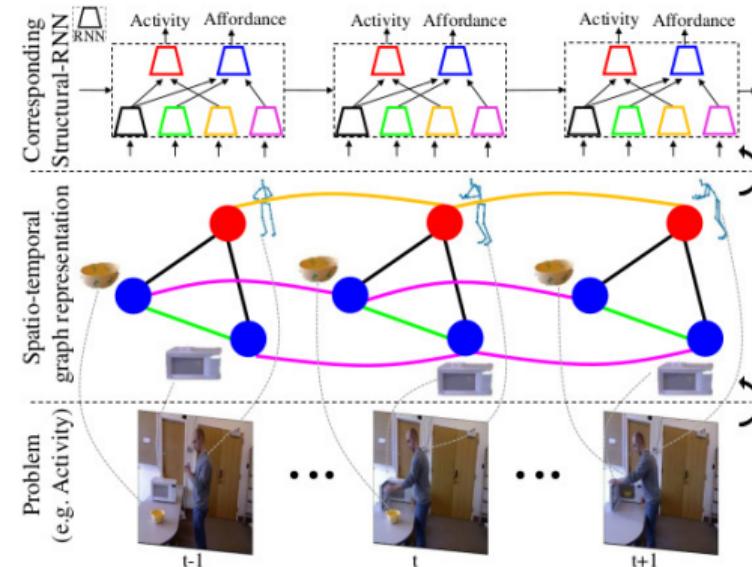
Convolutional GNN:

Different weights at different time-stamps

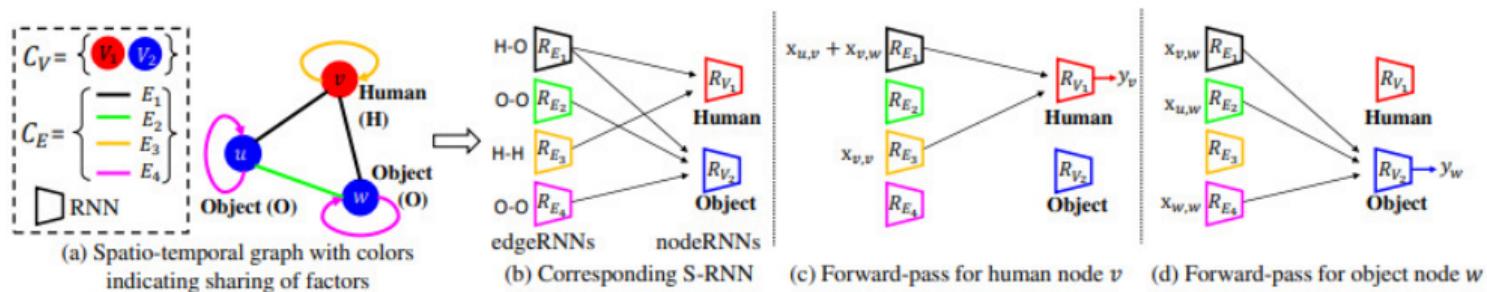
A compact review of ... GNN

Structured RNN (S-RNN)

GNN: Spatio-Temporal adaptation



Construction of S-RNN for a video shot



CVPR-16 Paper *, (Presentation Video) *

Applications

- Rec-GNN / Conv-GNN (still images - 2D/3D)
 - ▶ Fine-grained classification (e.g. bird species)
 - ▶ 3D point cloud processing (LiDAR)
- S-RNN (video / motion picture)
 - ▶ Human action recognition
 - ▶ Human/Robot - Object Interaction
 - ▶ Human Motion Modeling

Limitations and Future Research

- 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
- Model depth
 - ▶ Performance (accuracy) drops with depth

Quiz



Quiz 07-05

End of Module 07-05