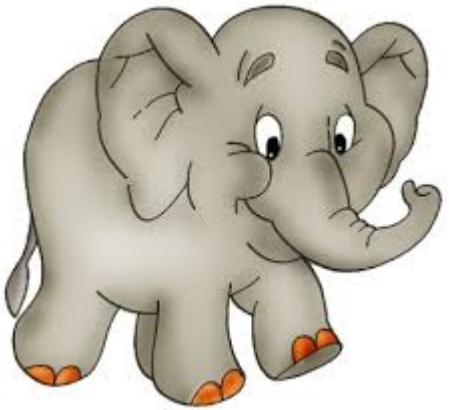
ELL 788
Computational Perception & Cognition

July – November 2015

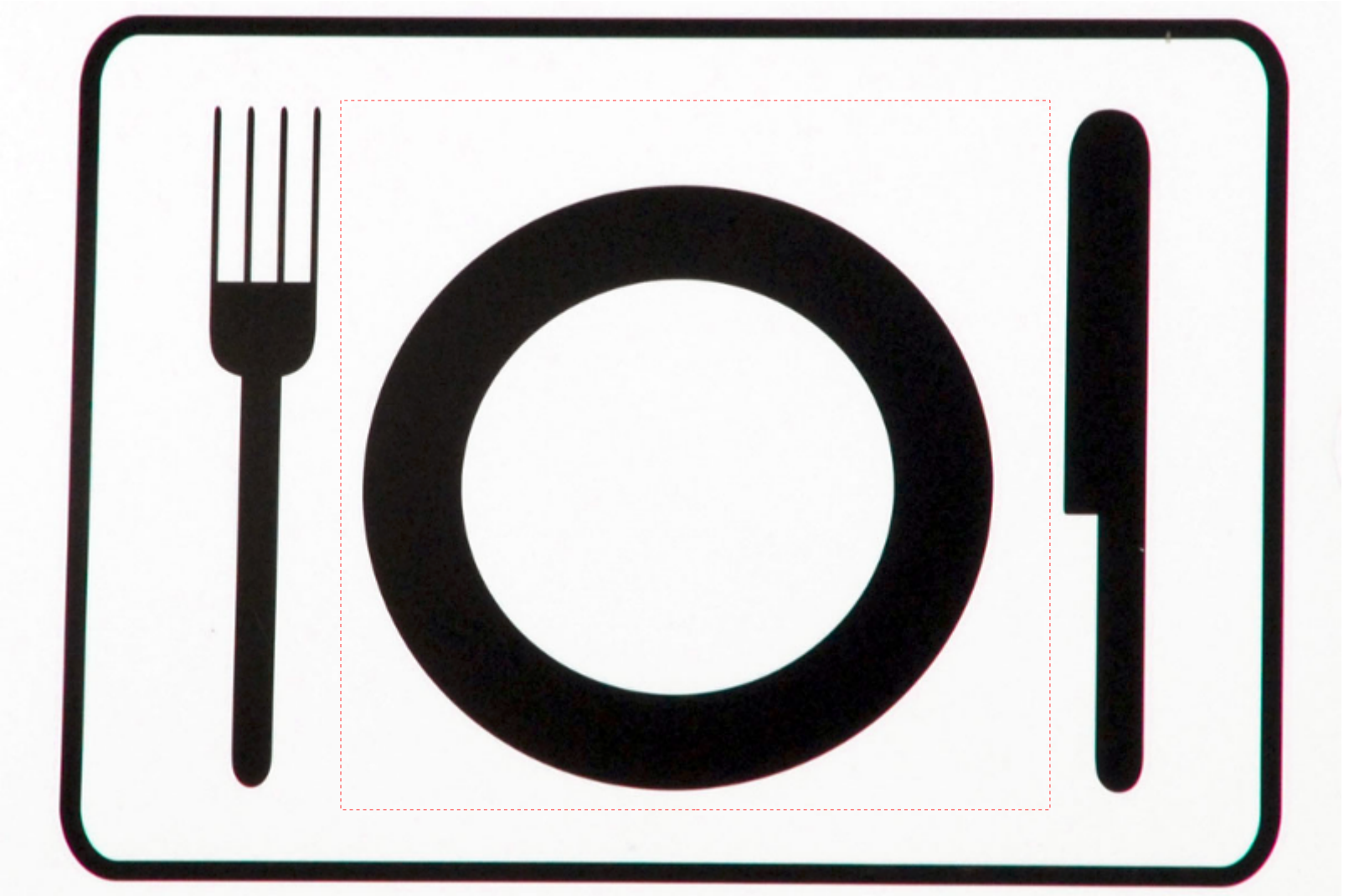
Module 6

Role of context in object detection

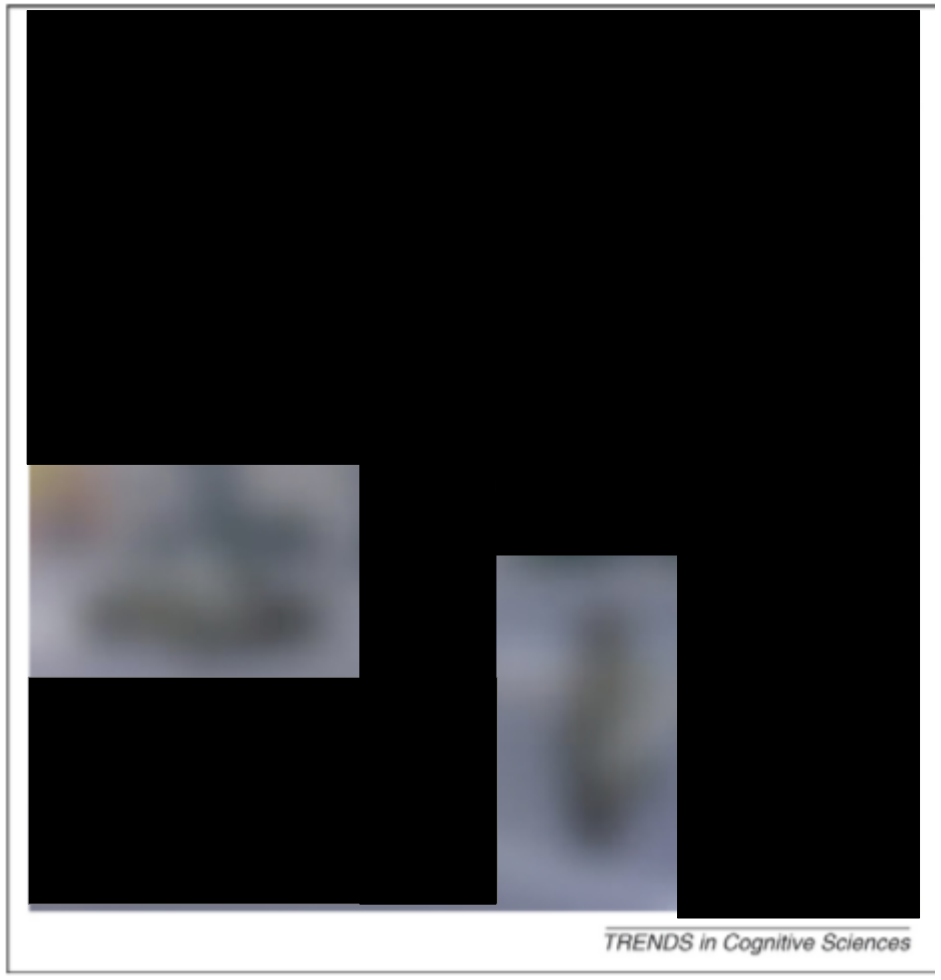
Objects and cognition

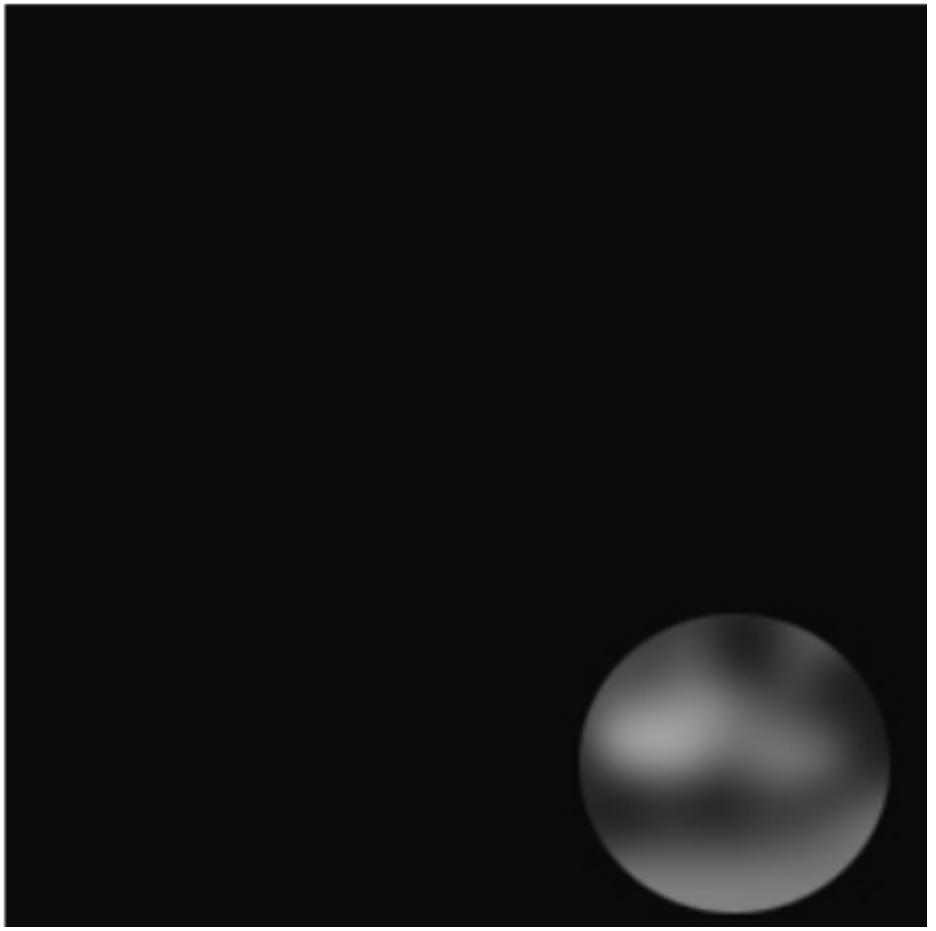


Ambiguous objects



Unfavorable viewing condition

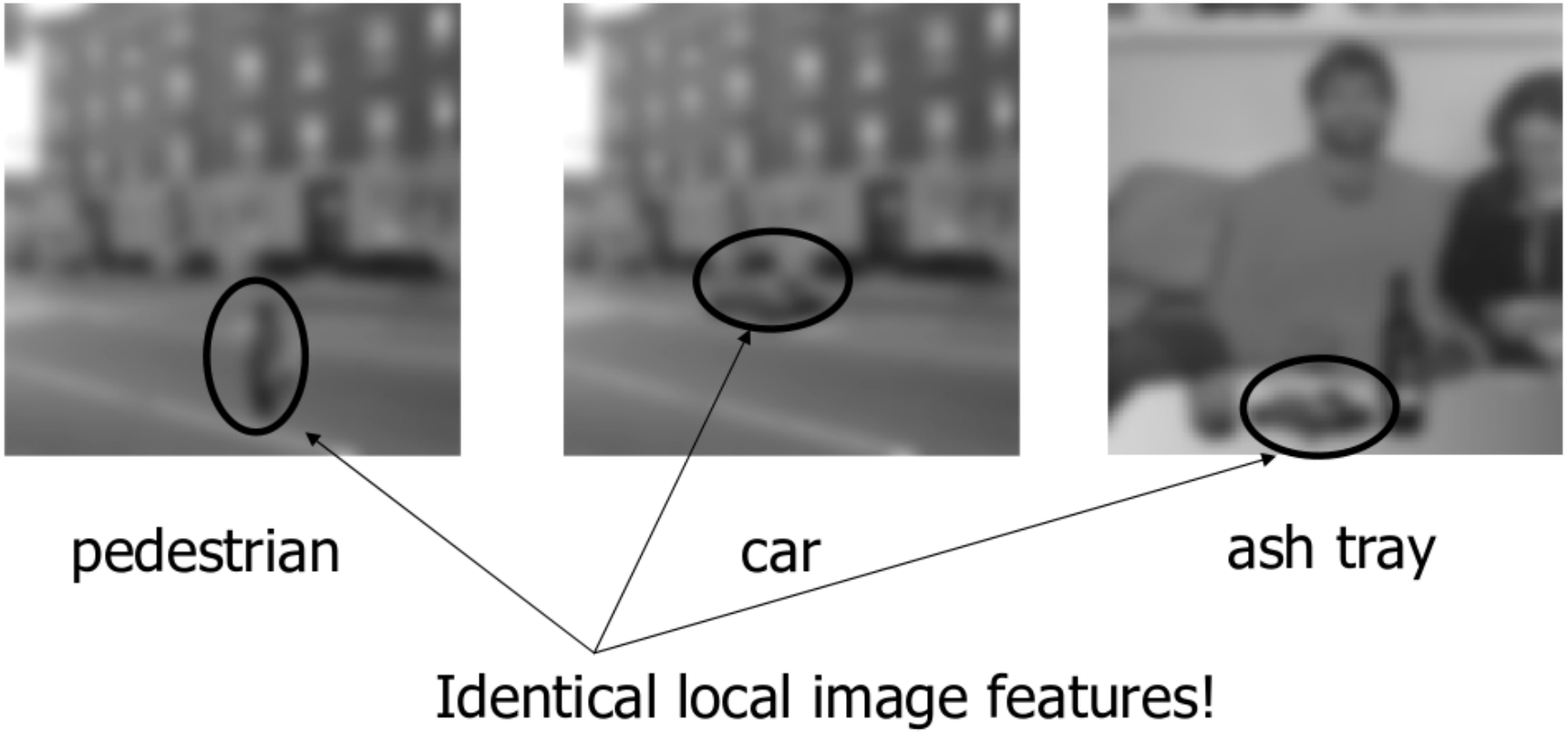




Context helps in object recognition

Source: Torrabla 2007

Context for resolving ambiguity

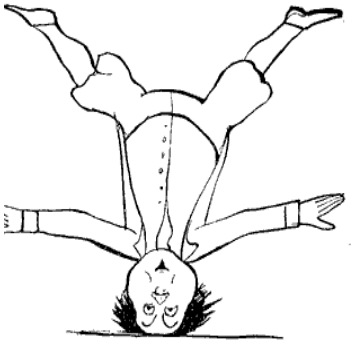


Source: http://www.cs.umd.edu/~djacobs/CMSC828/Context.pdf

Types of context

Semantic

Spatial arrangement



Pose

Summary

- Under favorable conditions
  - The multiplicity of cues (color, shape, texture) in the retinal image provides enough information to unambiguously determine the object category.
  - The object recognition mechanisms can rely exclusively on intrinsic object features.
  - Can robustly handle many transformations such as displacement, rotation, scaling, changes in illumination, etc
  - Context may not be required

- Under unfavorable conditions
  - Intrinsic object information alone cannot yield reliable results
  - When the object is immersed in its typical environment, recognition of the object becomes reliable
  - In real-world scenes, intrinsic object information is often degraded due to occlusions, illumination, shadows, peripheral vision and distance.
  - Context is necessary in order to build efficient and reliable algorithms for object recognition

Even when the object can be recognized with its intrinsic properties, context makes the recognition more efficient and reliable.

In context object recognition

Context: Prior knowledge about a closed world
(Location of the objects and structure of the scene)

Examples: Cars to be found on the road and pedestrians on the footpath
A building must be standing on the ground
Sky must be over the horizon

In-context object recognition exploits

- A set of intrinsic properties of the objects
- An object based description of context
- A set of rules, spacifying (relative) location of the objects in a scene

Logical reasoning *vs*. Evidential reasoning

Logical reasoning:

A $\rightarrow$ B: Observe A to infer B (sound)

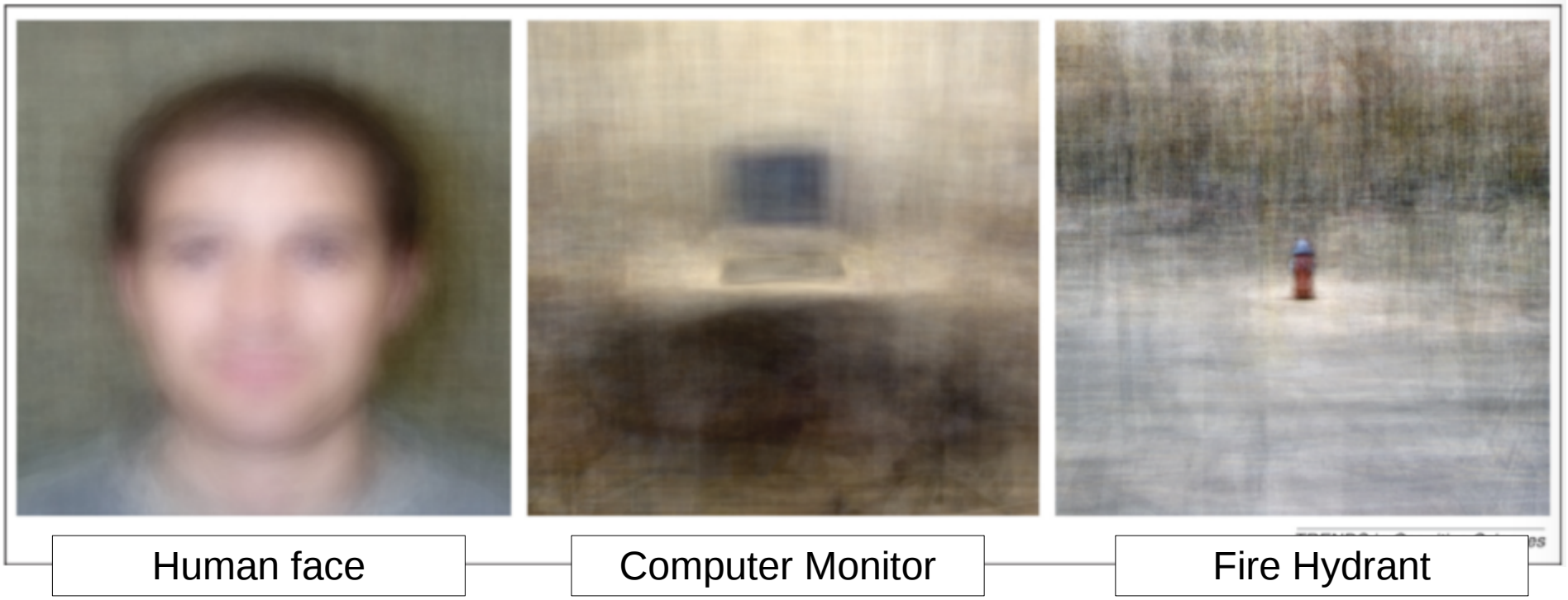
Evidential reasoning:

If A $\rightarrow$ B (A *causes* B), A is a plausible explanation of B
Observe B to infer A (weak)

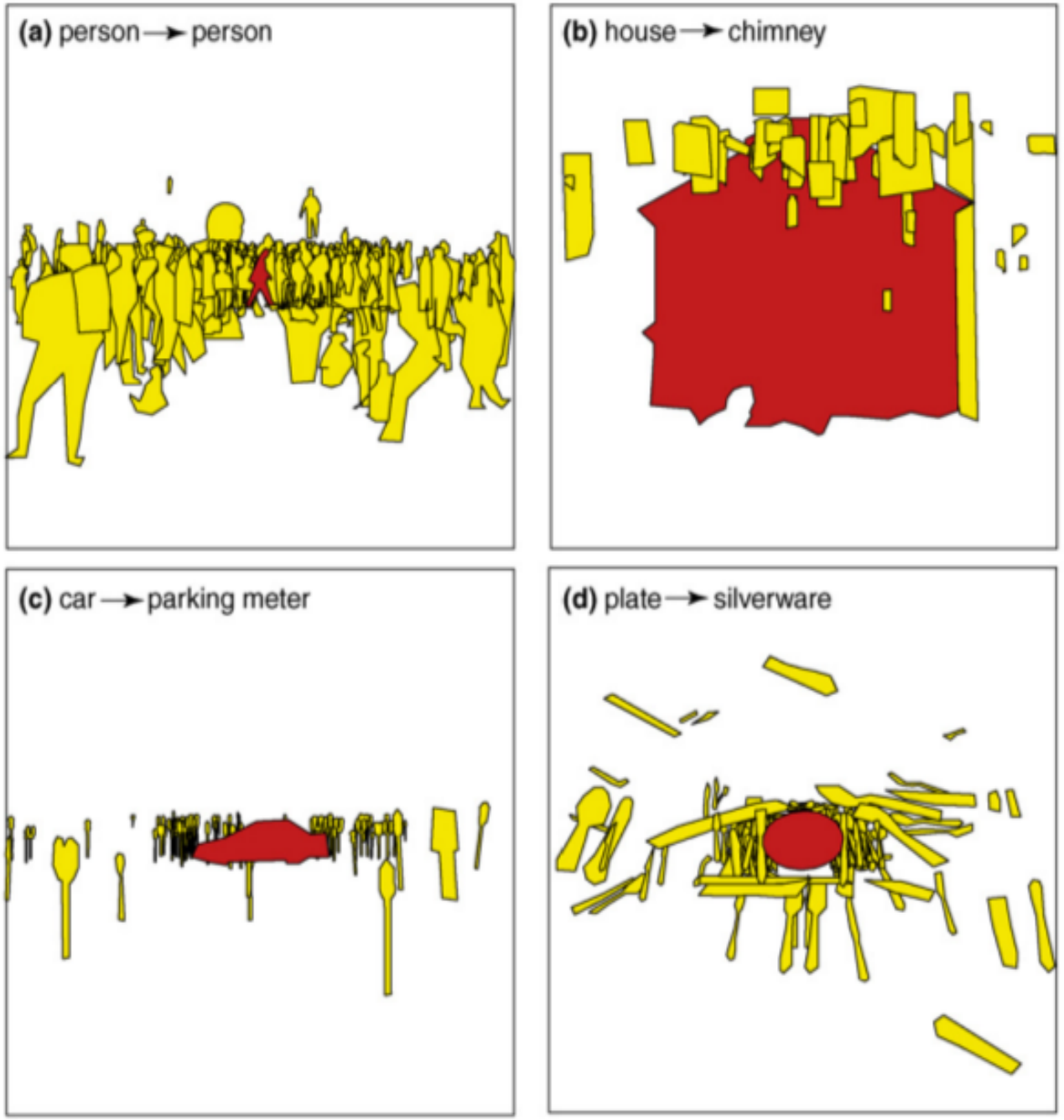
Rain causes **Wet Road** **Rain** is a plausible explanation of **Wet Road**
Observing wet road makes us to *believe* that it rained

- Belief is graded [0,1] – not binary {0,1} as in logical reasoning
- In general, more evidences result in more belief
- Robust against missing / erroneous observations

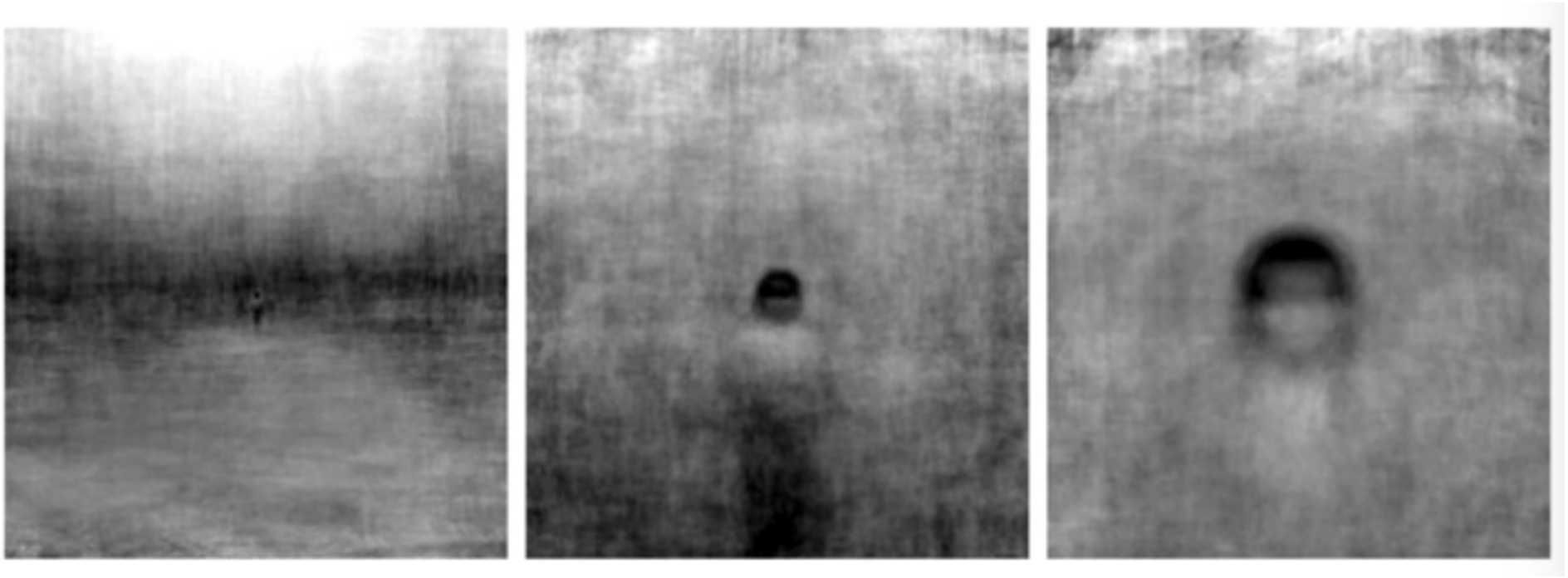
- Bayesian reasoning (Bayesian network) is a popular tool for evidential reasoning
- The “rules” are encoded as prior probabilities



Structure of objects and backgrounds



Dependencies amongst objects



Human head with different backgrounds in three different scales

(The background does not average out to be a uniform gray)

Background pixels provides information on

(1) Likelihood of finding an object
(2) Likelihood of places to find an object

Human perception system exploits such contextual information

Bayesian model for in-context object recognition

Let $O_1, O_2, \ldots O_N$ represent a set of visual objects in an image

- O = {o, **x**, σ, ...}: o = label (car, tree,etc.), **x** = location, σ = size, ...

Let $\boldsymbol{v}_1, \boldsymbol{v}_2. \ldots \boldsymbol{v}_N$ represent their visual properties

- some local measurements in that region, e.g. pixel intensities, color, texture, ... or a combination

Joint distribution of N objects and their visual properties (in a scene) is given by

$$P(O_1, \ldots, O_N, \mathbf{v}_1, \ldots, \mathbf{v}_N) \simeq \left[\prod_i^N P(\mathbf{v}_i \mid O_i)\right] P(O_1, \ldots, O_N)$$

- The simplification assumes conditional independence of visual properties
- The term $P(O_1, O_2 \ldots O_N)$ incorporates the contextual information

Approximation ... object recognition without context

Given an image: $$P(O \mid \mathbf{v}) = \frac{P(\mathbf{v} \mid O)}{P(\mathbf{v})} P(O)$$

- $\boldsymbol{v}$ is the image feature (computed over the complete image)
  - Vector of local measures – not global average
  - Large dimensionality of $\boldsymbol{v}$ makes it an ill-posed problem

- Approximation done in practical computer vision systems

$$P(O \mid \mathbf{v}) \simeq P(O \mid \mathbf{v}_L) = \frac{P(\mathbf{v}_L \mid O)}{P(\mathbf{v}_L)} P(O) \qquad v_L = v_{B(x, g(\sigma))}$$

- Simplification assumes that
  - an object is recognized by its local image features only
  - Image features of the background does not matter

Object detection in context

Image features are split into two sets: $\mathbf{v} = \left\{\mathbf{v}_{B(\mathbf{x},\epsilon)}, \mathbf{v}_{\bar{B}(\mathbf{x},\epsilon)}\right\} = \{\mathbf{v}_L, \mathbf{v}_C\}$

Local beighborhood and complementary locations (context)

$$P(O \mid \mathbf{v}) = P(O \mid \mathbf{v}_L, \mathbf{v}_C)$$

The object-centered approach approximates this to

$$P(O \mid \mathbf{v}_L, \mathbf{v}_C) \;=\; P(O \mid \mathbf{v}_L)$$

$$P(O \mid \mathbf{v}) = \frac{P(O, \mathbf{v})}{P(\mathbf{v})} = \frac{P(\mathbf{v}_L \mid O, \mathbf{v}_C)}{P(\mathbf{v}_L \mid \mathbf{v}_C)} P(O \mid \mathbf{v}_C)$$

Posterior distribution of local features when the object O is present in the context, normalized by distribution of local features in the context

Probability of occurence of the object in context (includes object attributes like class, location, size, ...)

Analyzing contextual priors $P(O|V_C)$

O represents object attributes: class, position, size, orientation ...

Let $O = \{o, \mathbf{x}, \sigma\}$

$$P(O \mid \mathbf{v}_C) = P(\sigma \mid \mathbf{x}, o, \mathbf{v}_C) P(\mathbf{x} \mid o, \mathbf{v}_C) P(o \mid \mathbf{v}_C)$$

Object Priming:
Probability of an object class given visual features of context

Focus of attention:
Probability of appearance of an object at a location in a certain context

Scale selection:
Probability of size of an object of a given class, when it appears at a location in a certain context

Other factorizations are also possible leading to diffent interpretations

A computational model of contextual object recognition

Object Priming

- What are the most likely objects to be found in the context
- Reduces the number of features needed for discriminating between those objects

Focus of attention

- What are the most likely places to find those objects
- Reduces search space

Scale selection

- What is the most likely scale for an object to appear at that place
- Reduces requirement of multi-scale seach

Benefits of this approach

- Reduces complexity
- Robust

Context representation

Context $\mathbf{v}_C$ has large dimensionality – need for dimensionality reduction

Subordinate scene description:
- Represent context as a collection of (other) objects in the scene
- Boils down to out-of-context object recognition

Holistic Context representation
- Spatial envelop representation – global or windowed (seen in last class)

Learning the priors

Learn over a large annotated data set
- Similar scenes – Object priming
- Similar objects in similar scenes – Focus of attention
- Similar locations for similar objects in similar scenes – Scale selection

Summary

- Context provides a strong cue for object recognition
- Strong correlation between statistical distribution of objects and environment
  - Differential distributions in different environments
- Holistic perception of a scene sets the context
  - Provides information on expected object types, their locations and their sizes
- Can be be used in practical computer vision systems to improve efficiency and robustness

References

- Torralba. Contextual Priming for Object Detection. 2003
- Oliva, Park and Konkale.
  Representing, perceiving, and remembering the shape of visual space (Book Chapter), 2010
- Oliva & Torralba. The role of context in object recognition, 2007