EECS 4422/5323 Computer Vision

Image Understanding 4

Calden Wloka

21 October, 2019

Announcements

- Midterm next week (October 28th) in class
- Sample Midterm questions online
- Remember to vote if you're able!
- Labs this week: Basic deep learning
- Assignment 1 marks out soon...

A Quick Review...

The *Associative* property:

$$(a+b)+c = a+(b+c)$$

The *Commutative* property:

$$a+b+c = a+c+b$$

Outline

- Feature Binding
- Border Ownership
- Visual Question Answering

Feature Relationships

The interactions between features is often of primary importance to understanding an image. Grouping features belonging to the same object or visual element is known as *feature binding*.

Feature Relationships

The interactions between features is often of primary importance to understanding an image. Grouping features belonging to the same object or visual element is known as *feature binding*.

Feature binding is a very general concept, and can refer to many different avenues of feature grouping, including:

- Spatial relationships

Feature Relationships

The interactions between features is often of primary importance to understanding an image. Grouping features belonging to the same object or visual element is known as *feature binding*.

Feature binding is a very general concept, and can refer to many different avenues of feature grouping, including:

- Spatial relationships
- Grouping across feature channels

Feature Relationships

The interactions between features is often of primary importance to understanding an image. Grouping features belonging to the same object or visual element is known as *feature binding*.

Feature binding is a very general concept, and can refer to many different avenues of feature grouping, including:

- Spatial relationships
- Grouping across feature channels
- Grouping by relative motion

A Motivating Example

We are often confronted with many objects which share common features, but are nevertheless distinct. Being able to identify which specific features belong to which object instance is often required for appropriately understanding and responding to an image.

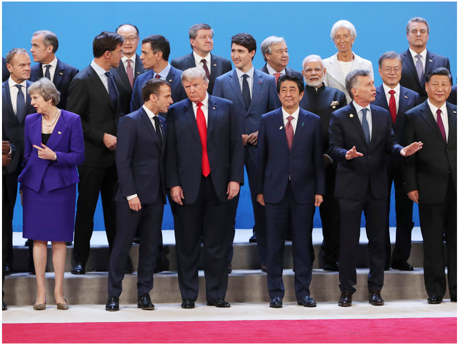


Image source: Gulf News

Relative Position Matters

It's not always sufficient for all the parts to be there, we also need those parts to be arranged correctly with respect to each other.

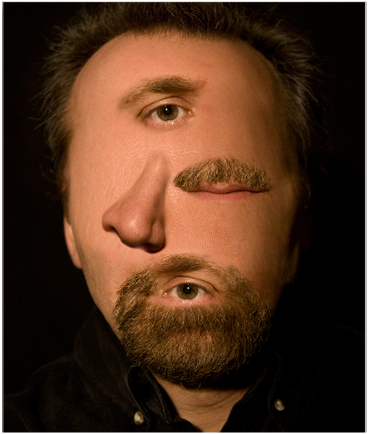


Image source: True Center Publishing

Likewise, Consistent Part Orientation Matters

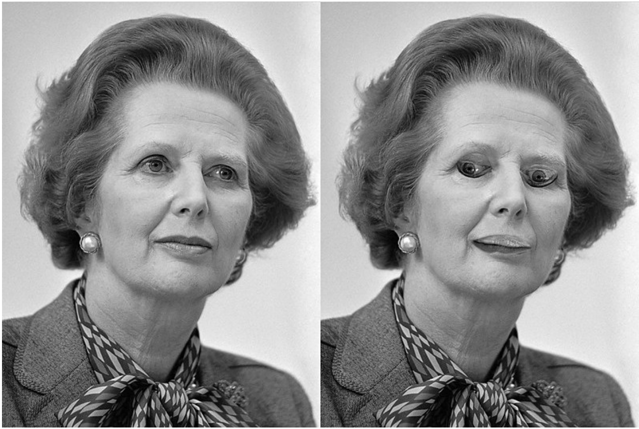


Image source: Thompson, 1980

The Relative Weighting of Parts Can Be Heuristically Inconsistent

The “Thatcher Illusion” displays how humans put different emphasis on part orientations based on the global orientation.

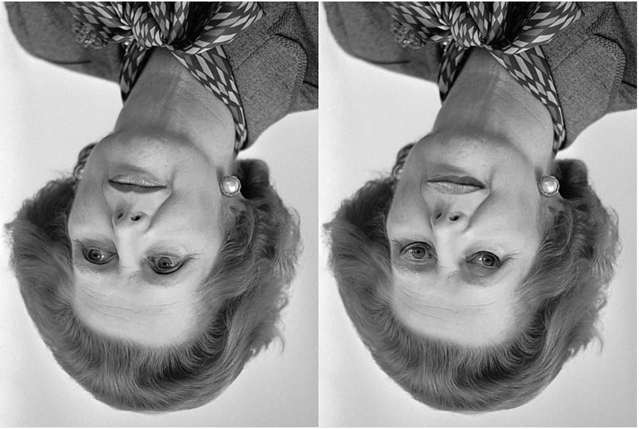


Image source: Thompson, 1980

Many Different Approaches to Spatial Encoding

- Constellation methods

Many Different Approaches to Spatial Encoding

- Constellation methods
- Implicit learned representation

Many Different Approaches to Spatial Encoding

- Constellation methods
- Implicit learned representation
- 3D models

Constellation Methods

Constellation methods encode spatial relationships in a graphical manner, assigning detected features or object components to nodes and then assigning edge values based on the relationship between the features.

Constellation Methods

Constellation methods encode spatial relationships in a graphical manner, assigning detected features or object components to nodes and then assigning edge values based on the relationship between the features.

Subgraphs which correspond to a valid object configuration are taken as valid detections.

Feature Topologies

Different arrangements of feature graphs can give rise to different behvaiours and computational complexities.

Fully connected models encode expected relationships for all parts, making them more robust but computationally intense.

"Star" models are contingent on a *landmark* feature against which all other features are relationally encoded.

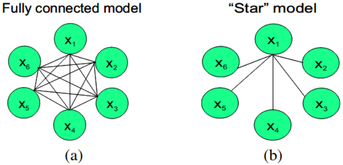


Image source: Fergus *et al.*, 2005

Parts-Based Example

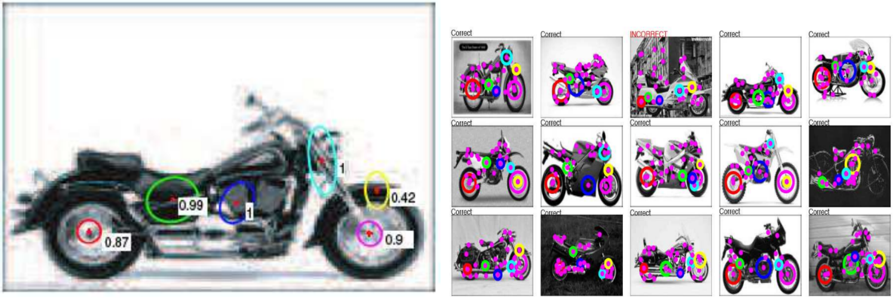


Image source: Fergus *et al.*, 2007

Challenges for Constellation Methods

- Occlusions

Challenges for Constellation Methods

- Occlusions
- Viewpoint changes

Challenges for Constellation Methods

- Occlusions
- Viewpoint changes
- Inter-class variation

Implicit Encodings

Rather than attempt to model the explicit spatial relationship between all sub-components or features of an object, deep networks implicitly represent the spatial relationships through a hierarchy of features.

Implicit Encodings

Rather than attempt to model the explicit spatial relationship between all sub-components or features of an object, deep networks implicitly represent the spatial relationships through a hierarchy of features.

With sufficient training data, we can learn a representation for multiple orientations and spatial scales.

Implicit Encodings

Rather than attempt to model the explicit spatial relationship between all sub-components or features of an object, deep networks implicitly represent the spatial relationships through a hierarchy of features.

With sufficient training data, we can learn a representation for multiple orientations and spatial scales.

Although all the different representations are related at output through the fully connected layers, their internal convolutional representations are largely disconnected, which can lead to unpredictable results.

A Major Challenge: Articulation and Deformation

As we have seen in panoramic image stitching, image transformations over rigid objects are typically *affine*, possibly subjected to a *projective warp*. This allows us to model spatial relationships as static values within some world coordinate space.

However, not all objects are rigid, and the relative spatial relationship between object components might vary over a range of possible positions. Internal object transformations typically falls under one of two classes: *articulated* movement and object *deformations*.

Articulation

Articulation refers to an object with rigid segments joined by mobile or flexible joints. Object transformations are typically constrained by a range of motion across these joints, though the number of possible configurations can grow rapidly.

3D Models

Although 3D are more complex to encode and parametrize, once learned they can be used to reconstruct any potential viewpoint or (in the case of an articulated model) object configuration. To encode the equivalent with 2D appearance based models, we would need to somehow discretize the view space for an object at sufficient resolution to interpolate across explicitly encoded viewpoints.

Implicit Encodings with Feedback

Although 3D models have not moved to the forefront of methods and constellation methods have fallen out of favour, more deep learning research is beginning to approach constellation style methods by incorporating feedback mechanisms to better enforce global constraints over the implicitly encoded spatial relationships of a feedforward network.

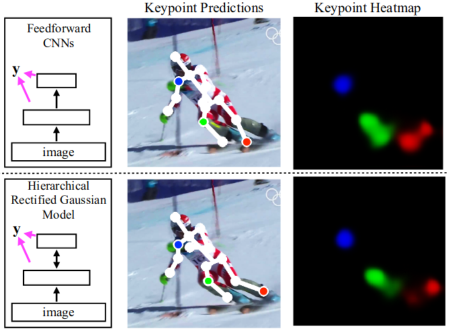


Image source: Hu & Ramanan, 2016

3D Models for Generative Applications

Where we do see explicit 3D models more commonly is in generative work. Modern approaches to these models are typically a hybrid of imposing model structure and constraints, but then learning the parameters of those constraints along with the image generator.

Image source: Holden *et al.*, 2016

Arbitrary Deformation

Objects for which a significant portion is made up of non-rigid material (*e.g*. flags, fluids, balloons, clouds) are particuarly challenging to characterize.

Arbitrary Deformation

Objects for which a significant portion is made up of non-rigid material (*e.g*. flags, fluids, balloons, clouds) are particuarly challenging to characterize.

These objects are frequently neglected in computer vision research, but are nevertheless potentially important targets. It is important to remember the existence of non-affine transformations when designing systems or evaluation the claims of others.

Sometimes spatial relationships aren't enough

Sometimes items can be spatially congruent, but still belong to different objects.

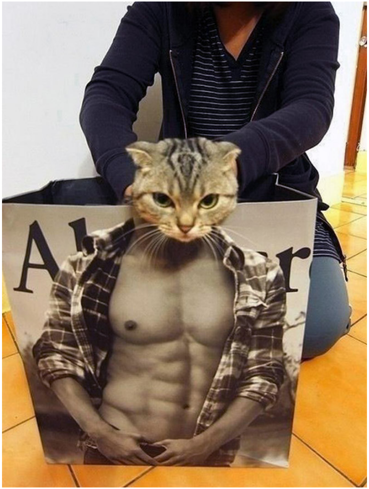


Image source: Sad and Useless

Detailing Object Attributes

In addition to disambiguating object occlusions, cross-channel feature binding is necessary for cohesively returning a set of object attributes.

Objects on the right could be described by:

- Surface reflectance
- Colour
- Size
- Shape

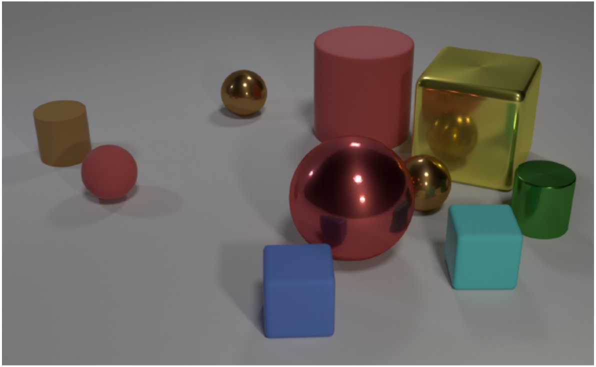


Image source: Johnson *et al.*, 2017

Constraints and Dynamic Binding

As mentioned several times in this course, the more you can constrain your problem to know what features you need to access for any given object, the better.

Constraints and Dynamic Binding

As mentioned several times in this course, the more you can constrain your problem to know what features you need to access for any given object, the better.

If you don't know *a priori* what features you need, then you either need to exhaustively compute object attributes across the image (computationally intensive and potentially intractable), or you need a method to dynamically determine which features are necessary (attention).

Occlusions and Crowded Fields

A crowded field of highly similar objects can be particularly challenging to disambiguate. One approach which is possibly useful is *border ownership*; the task of determining for a given line if it is an object boundary and, if so, which side of the boundary is "owned" by a given object.

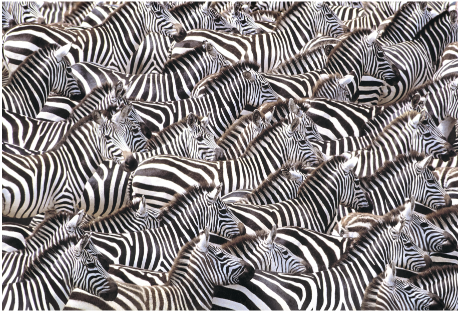


Image source: Original source unknown

A Brief Note on Occlusions

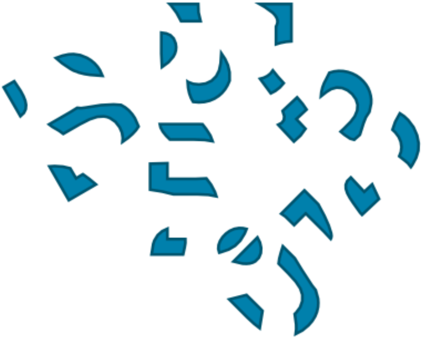


Image source: Bregman, 1981

A Brief Note on Occlusions

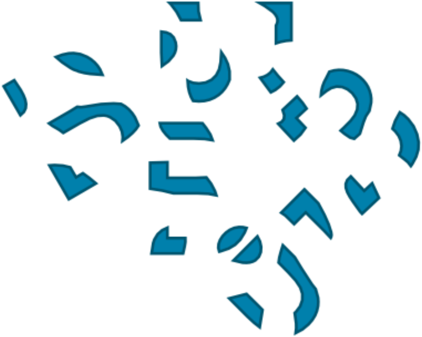


Image source: Bregman, 1981

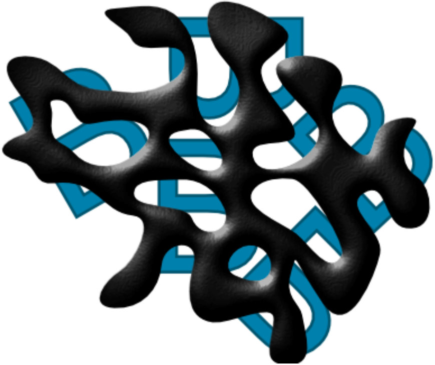


Image source: Bregman, 1981

Border Ownership and Semantic Segmentation

Border ownership shares superficial resemblance to the task of *semantic segmentation*, in which pixels are densely assigned class labels. However, there are some differences, such as internal border retention.

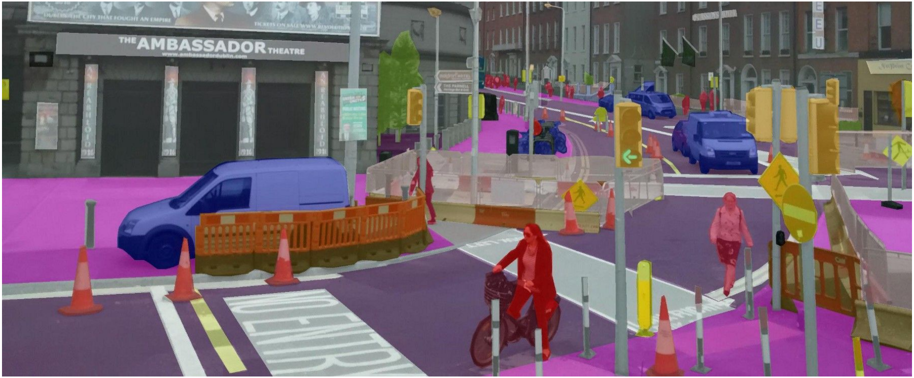


Image source: Nanonets

Different Types of Borders

- Point A shows a straightforward border between the koala (foreground) and tree (background)
- Point B shows self-occlusion of the koala with itself
- Point C shows another border between koala and tree, but with reversed spatial relationship

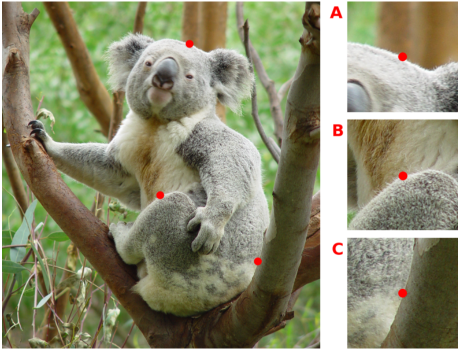


Image source: Williford & von der Heydt, 2013

General Purpose Border Ownership

General purpose border ownership is an open problem in computer vision; we can potentially take inspiration from biological recordings of neurons which respond to borders.

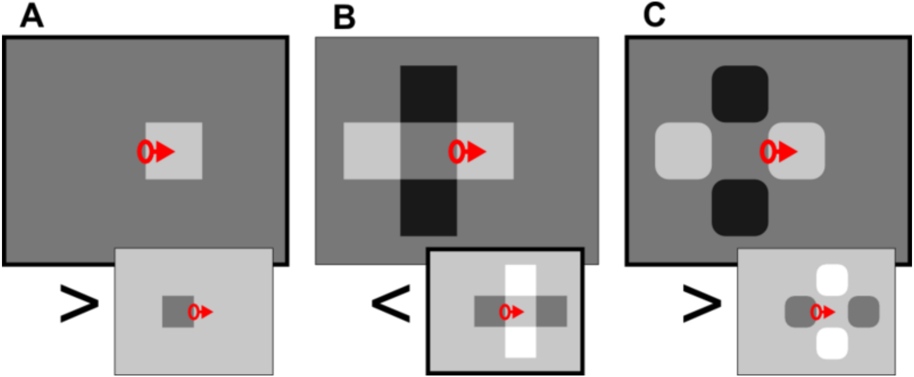


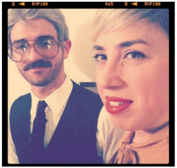
Image source: Qui & von der Heydt, 2007

Visual Question Answering

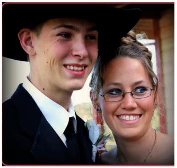
An emerging research area for which the types of questions raised in this lecture are of paramount importance is *Visual Question Answering (VQA)*.

Who is wearing glasses?

man

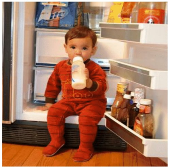


woman

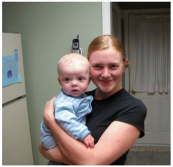


Where is the child sitting?

fridge

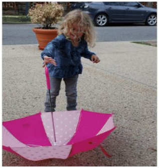


arms

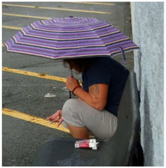


Is the umbrella upside down?

yes

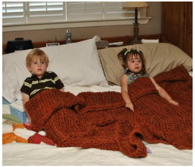


no

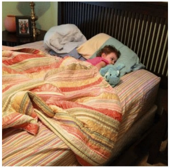


How many children are in the bed?

2



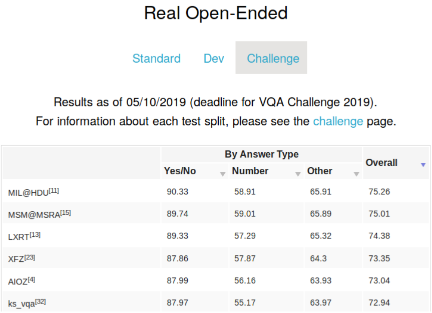
1



VQA Involves Other Modalities

A number of modalities involved in VQA research fall outside of the domain of this course, such as linguistic understanding. Nevertheless, the visual side of the problem area are highly emblematic of the open challenges which remain in visual understanding.

As we can see, the type of question being asked greatly impacts the accuracies which can be achieved.



Real Open-Ended

Results as of 05/10/2019 (deadline for VQA Challenge 2019).
For information about each test split, please see the challenge page.

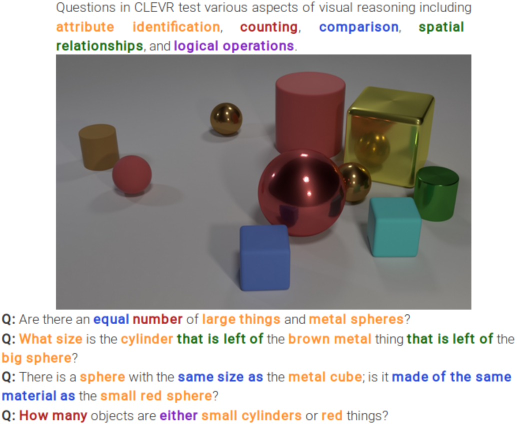
	By Answer Type			Overall
	Yes/No	Number	Other	
MIL@HDU[11]	90.33	58.91	65.91	75.26
MSM@MSRA[15]	89.74	59.01	65.89	75.01
LXRT[13]	89.33	57.29	65.32	74.38
XFZ[23]	87.86	57.87	64.3	73.35
AIOZ[4]	87.99	56.16	63.93	73.04
ks_vqa[32]	87.97	55.17	63.97	72.94

Image source: VQA Challenge 2019

CLEVR: An Alternative Approach to VQA Data

CLEVR is a generational framework for rendering block-world stimuli and corresponding questions.

How does this approach compare to the VQA Challenge from the previous slide?



Questions in CLEVR test various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.

Q: Are there an **equal number** of **large things** and **metal spheres**?

Q: **What size** is the **cylinder that is left of** the **brown metal** thing **that is left of** the **big sphere**?

Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?

Q: **How many** objects are **either small cylinders** or **red** things?

Image source: Johnson *et al.*, 2017