Topics in AI (CPSC 532S):
Multimodal Learning with Vision, Language and Sound

Lecture 4: Introduction to Computer Vision
Computer vs. human vision

Human Vision

*slide from V. Ordonex*
Computer vs. human vision

objects, scenes, people

Human Vision

*slide from V. Ordonex
Computer vs. human vision

 objects, scenes, people

Human Vision

Computer Vision

matrix of numbers

*slide from V. Ordonex
Computer vs. human vision

Human Vision

objects, scenes, people

Computer Vision

tensor of numbers

*slide from V. Ordonex
Computer Vision

Computer vision studies the **tools and theories** that enable the design of machines that can **extract useful information from imagery data** (images and videos) toward the goal of **interpreting the world**

*courtesy of Peter Meer*
Vision is Amazing Feat of Natural Intelligence

~ 55% of cerebral cortex in humans (13 billion neurons) are devoted to vision

more human brain devoted to vision than anything else
Challenges: Viewpoint invariance

Michelangelo 1475-1564

*slide credit Fei-Fei, Fergus & Torralba
Challenges: Lighting

*Image credit J. Koenderink*
Challenges: Scale

*slide credit Fei-Fei, Fergus & Torralba
Challenges: Deformation

*image credit Peter Meer*
Challenges: Occlusions

Rene Magritte 1965
Challenges: Background clutter

Kilmeny Niland 1995
Challenges: Local ambiguity and context

*image credit Fergus & Torralba*
Challenges: Local ambiguity and context

*image credit Fergus & Torralba*
Challenges: Motion

*image credit Peter Meer*
Challenges: Object inter-class variation

*slide credit Fei-Fei, Fergus & Torralba
Human vision …

Simple cells:
Response to light orientation

Complex cells:
Response to light orientation and movement

Hypercomplex cells:
Response to movement with end point

Hubel & Wiesel, 1959

Electrical signal from brain

Stimulus

No response
Response (end point)

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Computer vision ... the beginning ...

Larry Roberts

**Blocks World.** first thesis in computer vision, 1963

"the perception of **solid objects** is a process which can be based on the **properties of three-dimensional** transformations and the **laws of nature**"
Computer vision ... the beginning ...

Blocks World. first thesis in computer vision, 1963

Larry Roberts

"the perception of solid objects is a process which can be based on the properties of three-dimensional transformations and the laws of nature"

Figure 1. (a) A line drawing provides information only about the x, y coordinates of points lying along the object contours. (b) The human visual system is usually able to reconstruct an object in three dimensions given only a single 2D projection. (c) Any planar line-drawing is geometrically consistent with infinitely many 3D structures.

[ Since & Adelson, 1993 ]
Computer vision … the beginning …

**Static Equilibrium:** Forces and torques acting on a block should cancel each other out.

**Support Force Constraint:** Supporting object should have enough strength to provide contact reactionary forces.

**Volumetric Constraints:** All objects in the world must have finite volume & cannot penetrate each other.

[ Gupta, Efros & Hebert, 2010 ]
In 1966, Marvin Minsky at MIT asked his undergraduate student Gerald Jay Sussman to “spend the summer linking a camera to a computer and getting the computer to describe what it saw”

[Szeliski 2009, Computer Vision]
David Marr, 1970s

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, cs231n Stanford
David Marr, 1970s

Input Image

Edge Image

2 1/2-D sketch

3-D model

Input Image

Primal Sketch

Perceived intensities

Zero crossings, blobs, edges, bars, ends, virtual lines, groups, curves boundaries

2 1/2-D Sketch

Local surface orientation and discontinuities in depth and in surface orientation

3-D Model Representation

3-D models hierarchically organized in terms of surface and volumetric primitives

[ Stages of Visual Representation, David Marr ]

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Edges

<table>
<thead>
<tr>
<th>1</th>
<th>0</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>

A color picture of a steam engine

The Sobel operator applied to that image

*content from V. Ordonex*
David Marr, 1970s

**Input Image**

- Perceived intensities

**Edge Image**

- Zero crossings, blobs, edges, bars, ends, virtual lines, groups, curves boundaries

**2 1/2-D Sketch**

- Local surface orientation and discontinuities in depth and in surface orientation

**3-D Model Representation**

- 3-D models hierarchically organized in terms of surface and volumetric primitives

---

[Stages of Visual Representation, David Marr]

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Segmentation - GraphCuts

[ Shi & Malik, 2000 ]
David Marr, 1970s

Input image
- Input Image
- Perceived intensities

Edge image
- Primal Sketch
- Zero crossings, blobs, edges, bars, ends, virtual lines, groups, curves, boundaries

2 ½-D sketch
- 2 ½-D Sketch
- Local surface orientation and discontinuities in depth and in surface orientation

3-D model
- 3-D Model Representation
- 3-D models hierarchically organized in terms of surface and volumetric primitives

[ Stages of Visual Representation, David Marr ]

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Part-based Models

Generalized Cylinders

[ Brooks & Binford, 1979 ]

Pictorial Structures

[ Fischler & Elschlager, 1973 ]
Part-based Models

[ Sigal et al. 2004]

Monty Python’s Ministry of Silly Walks
Part-based Models

Monty Python’s Ministry of Silly Walks

[ Sigal et al. 2004]
David Marr, 1970s

Stages of Visual Representation, David Marr

- Input Image
  - Perceived intensities
- Primal Sketch
  - Zero crossings, blobs, edges, bars, ends, virtual lines, groups, curves boundaries
- 2 ½-D Sketch
  - Local surface orientation and discontinuities in depth and in surface orientation
- 3-D Model Representation
  - 3-D models hierarchically organized in terms of surface and volumetric primitives

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Face Detection 1999-2000

[ Viola & Jones, 2001 ]
Feature-based Vision

[ David Lowe, 1999 ]
Image content is transformed into local feature coordinates that are \textbf{invariant}
to translation, rotation, scale and imaging parameters

[ David Lowe, 1999 ]
SIFT Descriptor

[David Lowe, 1999]
Massive 3D Reconstructions

[ Agarwal, Furukawa, Snavely, Curless, Seitz, Szeliski, 2010 ]
Massive 3D Reconstructions

[ Agarwal, Furukawa, Snavely, Curless, Seitz, Szeliski, 2010 ]
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that arrive from our eyes. For a long time it was thought that the retinal image was transmitted point by point to visual centers in the brain; the cerebral cortex was a movie screen, so to speak, upon which the image in the eye was projected. Through the discoveries of Hubel and Wiesel we now know that behind the origin of visual perception in the brain there is a considerably more complicated course of events. By following the visual impulses along their path to the various cell layers of the optical cortex, Hubel and Wiesel have been able to demonstrate that the message about the image falling on the retina undergoes a step-wise analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.

---

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% jump in exports to $750bn, compared with a 18% rise in imports to $660bn. The figures are likely to further annoy the US, which has long argued that China's exports are unfairly helped by a deliberately undervalued yuan. Beijing agrees the surplus is too high, but says the yuan is only one factor. Bank of China governor Zhou Xiaochuan said the country also needed to do more to boost domestic demand so more goods stayed within the country. China increased the value of the yuan against the dollar by 2.1% in July and permitted it to trade within a narrow band, but the US wants the yuan to be allowed to trade freely. However, Beijing has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.
Bag-of-Visual-Words

*slide credit Li Fei-Fei
Beyond Bag of Features

[ Lazebnik, Schmid, Ponce, 2006 ]
Deformable Part Models

[ Felzenswalb, McAllester, Ramanan, 2009 ]
Deformable Part Models

[ Felzenswalb, McAllester, Ramanan, 2009 ]
Hierarchical Models

[Fidler, Leonardis, CVPR 2007]
PASCAL Visual Object Challenge (VOC)

[ Everingham et al. 2006-2012 ]
Effectiveness of **Data**

[ Hays, Efros, ACM Siggraph 2007 ]

[ Hays, Efros, CVPR 2008 ]
ImageNet Benchmark

22K categories and 14M images

- Animals
  - Bird
  - Fish
  - Mammal
  - Invertebrate
- Plants
  - Tree
  - Flower
  - Food
  - Materials
- Structures
  - Artifact
  - Tools
  - Appliances
  - Structures
- Person
- Scenes
  - Indoor
  - Geological Formations
  - Sport Activities

Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009
AlexNet on ImageNet

* image from CV-Tricks.com

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kri@cse.utoronto.ca
Ilya Sutskever
University of Toronto
ilya@cse.utoronto.ca
Geoffrey E. Hinton
University of Toronto
hinton@cse.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 7.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 60,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called “dropout” that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

Model | Top-1 (val) | Top-5 (val) | Top-5 (test)
--- | --- | --- | ---
SIFT + FVs [7] | — | — | 26.2%
1 CNN | 40.7% | 18.2% | —
5 CNNs | 38.1% | 16.4% | 16.4%
1 CNN* | 39.0% | 16.6% | —
7 CNNs* | 36.7% | 15.4% | 15.3%

[Krizhevsky, Sutskever, Hinton, NIPS 2012]
Success of Deep Learning

The Image Classification Challenge:
1,000 object classes
1,431,167 images

ImageNet Large Scale Visual Recognition Challenge

Russakovsky et al. arXiv, 2014
Final thought ...

- Model based, compositional, primitives, inverse graphics
- Hand-crafted features for given invariances & matching
- Hand-crafted features with learned statistical models on top
- Joint learning of features and statistical models for recognition
CapsuleNET

Going **back to inverse** graphics

*image credit medium.com*
Going **back to inverse** graphics

[ Sabour, Frosst, Hinton, NIPS 2017 ]

*image credit medium.com*
Going back to inverse graphics

[ Sabour, Frosst, Hinton, NIPS 2017 ]
Neural Modular Networks

[ Andreas, Rohrbach, Darrell, Klein, CVPR 2016 ]