Visual Perception – Part II

CPSC 470 – Artificial Intelligence
Brian Scassellati

2D Convolution Example

Starting Image

18	64	32	10	9	14	14
10	20	40	60	20	40	10
39	56	24	25	83	20	55
23	57	85	94	39	5	60
23	64	46	83	7	24	73
52	35	31	55	63	35	92
48	56	83	65	93	20	11

$\circledast$

Convolution Kernel

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

=

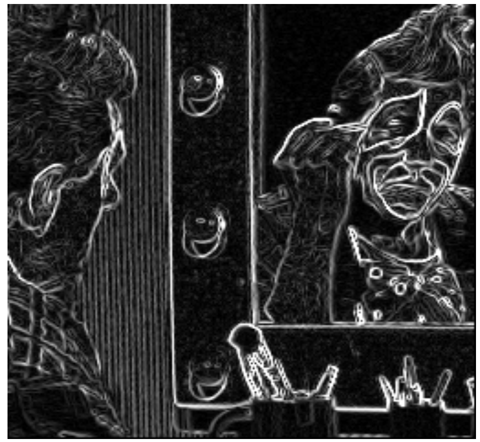
Resulting Image

	59					

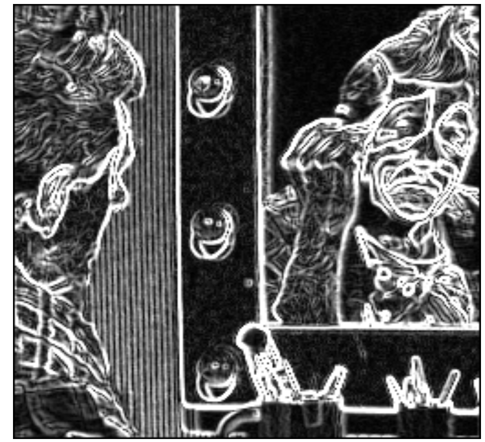
-1*18 + 0*64 + 1*32
-2*10 + 0*20 + 2*40
-1*39 + 0*56 + 1*24 = 59

Comparison of Edge Detectors

Original image



Results using Roberts Cross



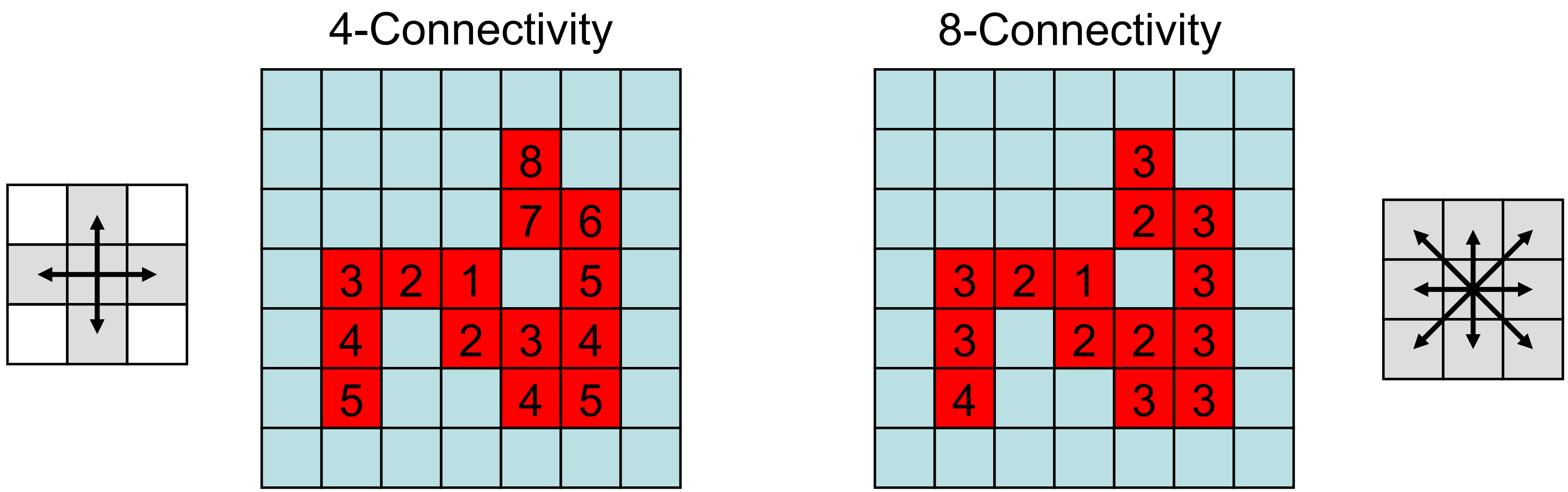
Results using Sobel

Zero crossings with σ = 2.0

Canny with σ = 1.0, T1 = 255, T2 = 1

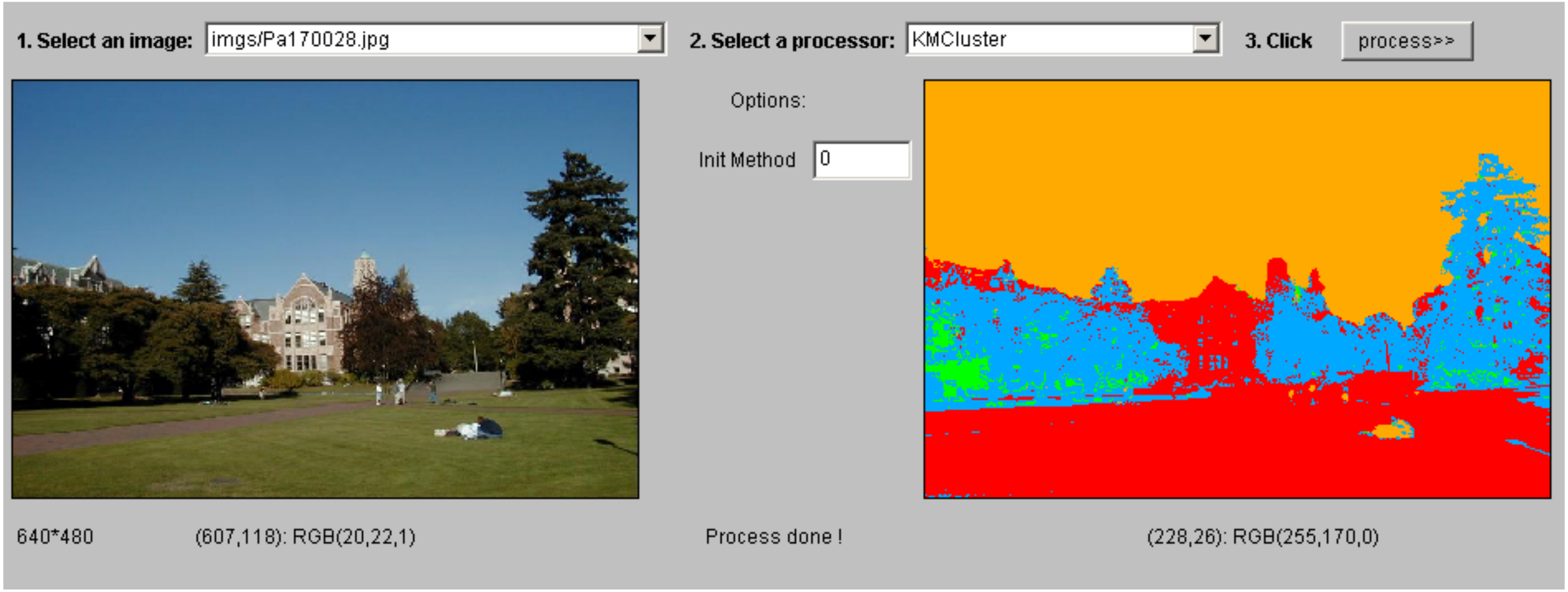
Canny with σ = 2.0, T1 = 128, T2 = 1

Segmentation via Region Growing



- Region growing techniques start with one pixel of a potential region and try to grow it by adding adjacent pixels until the pixels being compared are too dissimilar.
- The first pixel selected can be just the first unlabeled pixel in the image or a set of seed pixels can be chosen from the image.
  - Usually a statistical test is used to decide which pixels can be added to a region.

K-Means Example 1

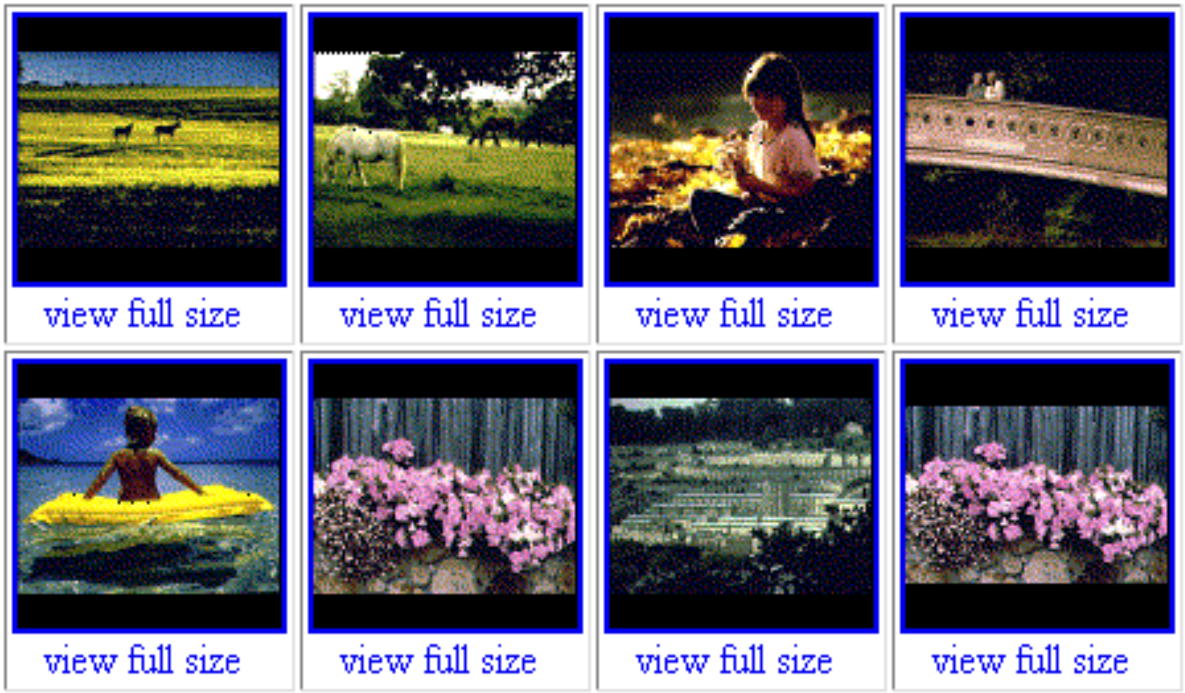


Today: Other Vision Tasks

- Finding similar images
- Extracting 3D structure
- Object recognition

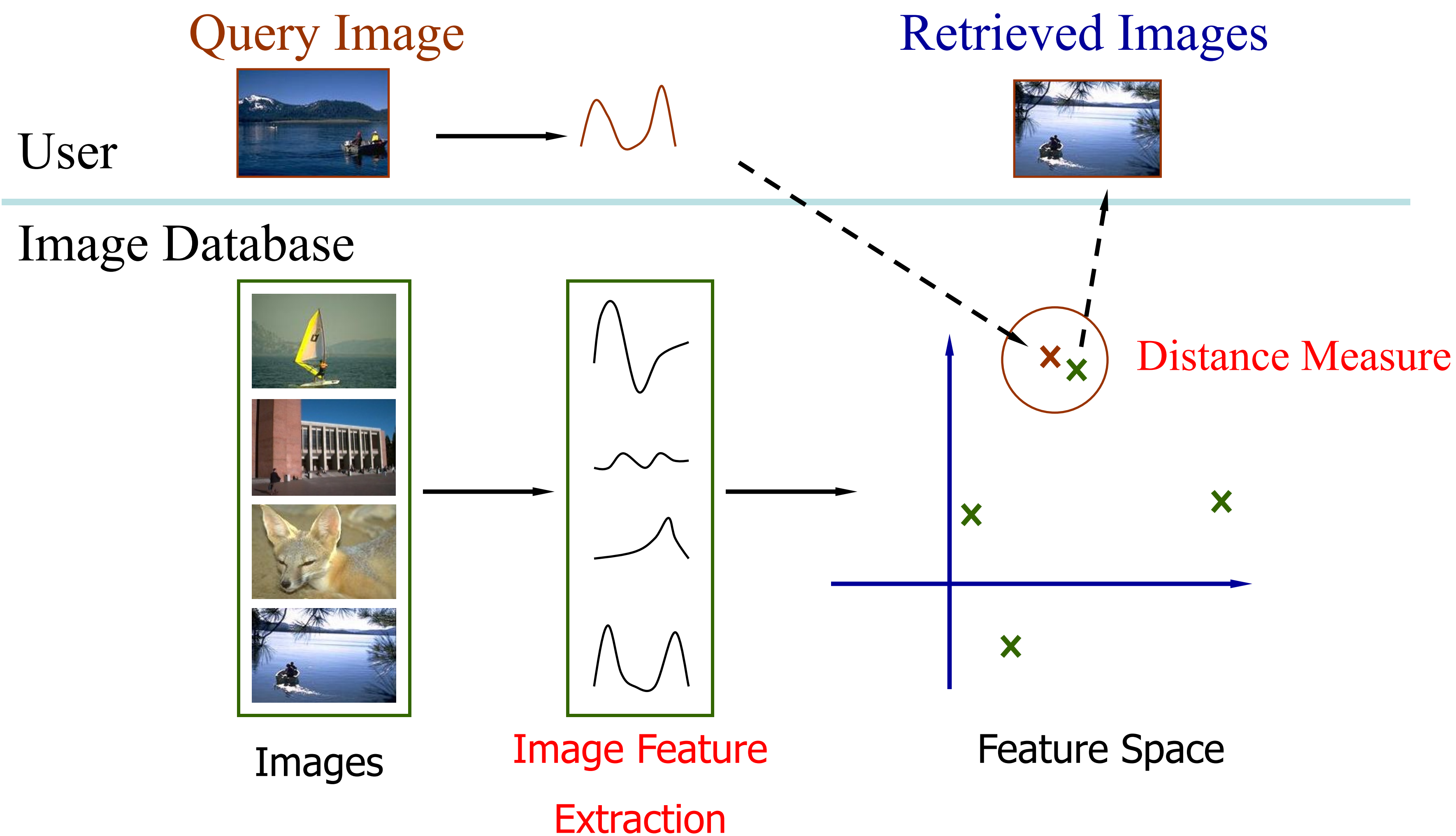
Finding Similar Images

Finding Similar Images



- Uses
  - Identification/recognition
  - Web-based searches
  - Databases
  - Medical imaging

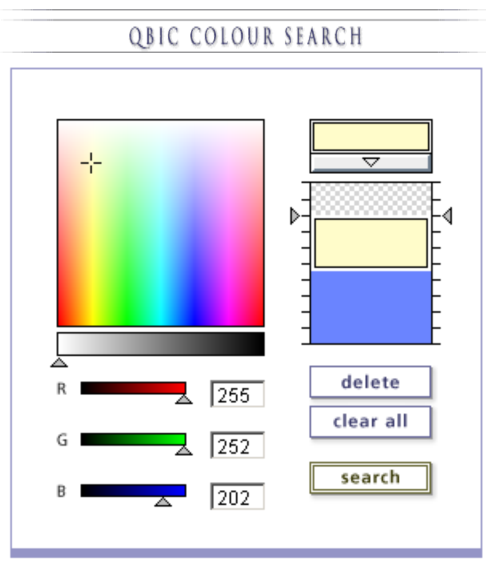
Searching for Image Similarity

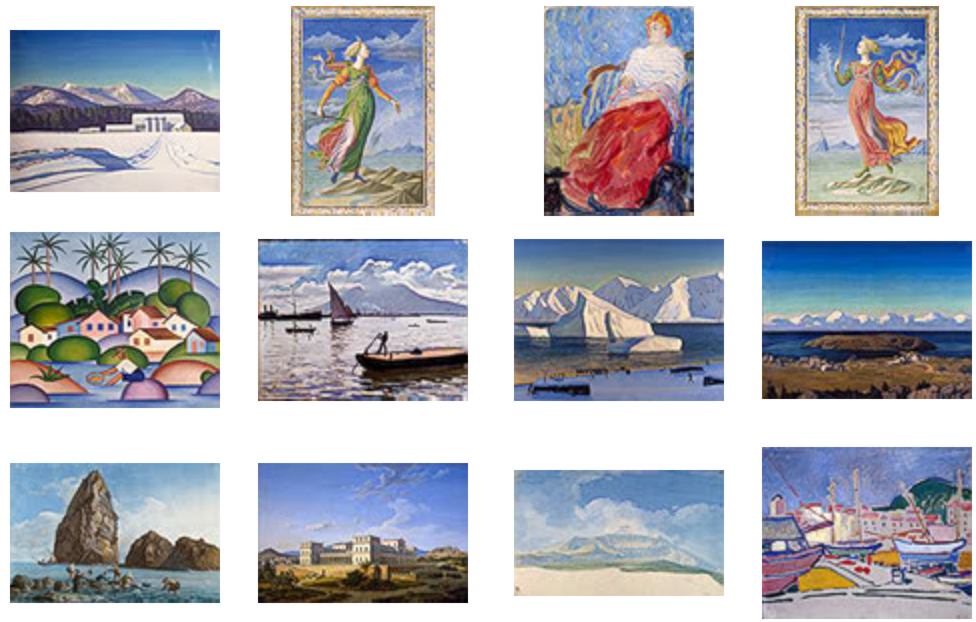


Sample Visual Features

- Color (histograms, gridded layout, wavelets)
- Texture (Laws, Gabor filters, local binary partition)
- Shape (first segment the image, then use statistical or structural shape similarity measures)
- Objects and their Relationships
  - This is the most powerful, but you have to be able to recognize the objects!

IBM's Query by Image Content (QBIC) at the Hermitage Museum



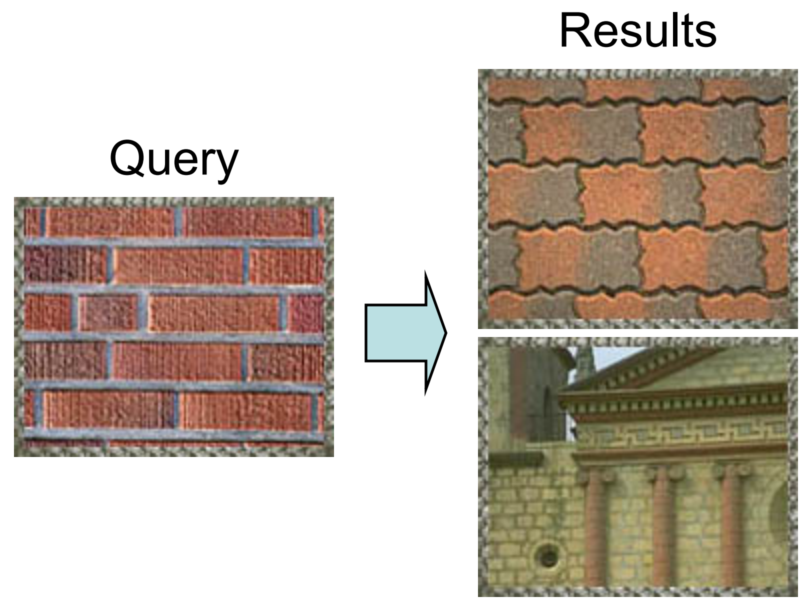


- Query based upon color histogram distribution
- In use as on many on-line image database sites (museums, clip art, etc.)

IBM's Query by Image Content (QBIC) at the Hermitage Museum

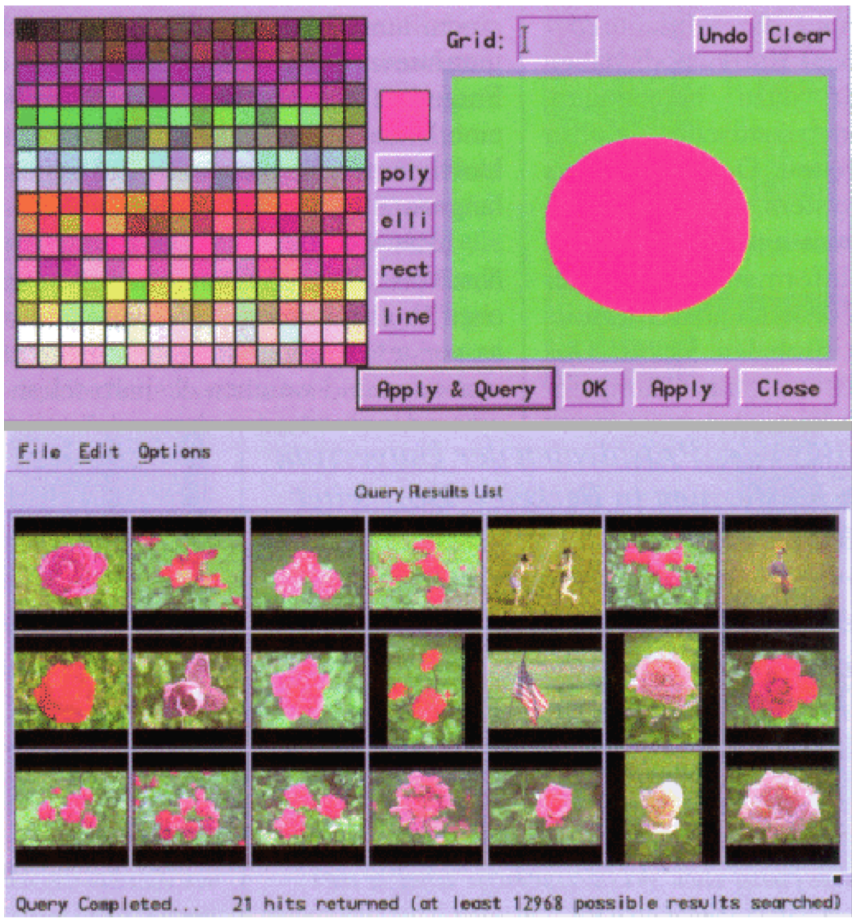
- Query based upon color histogram distribution
- In use on many on-line image database sites (museums, clip art, etc.)

QBIC texture



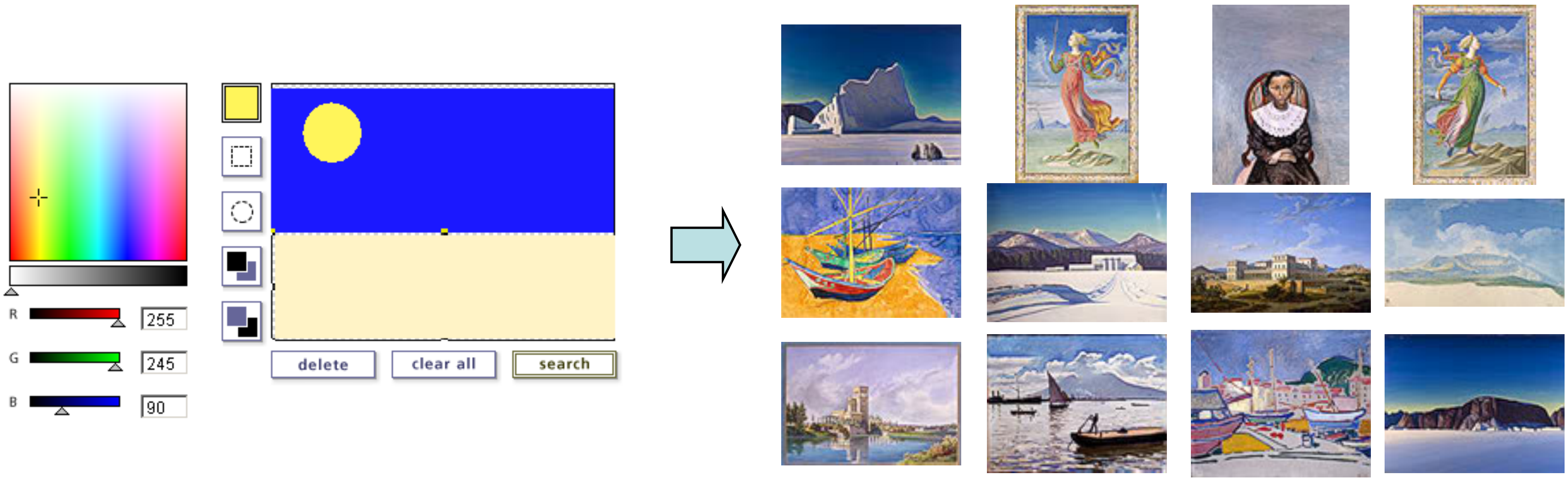
- What does it mean to have similarity of texture?
- Textels: spatial patterns (defined by a set of filters) that are repetitive across a region
- Good results on simple, well-defined textures
- Poor results on complex scenes

Color and Shape (really spatial distribution)



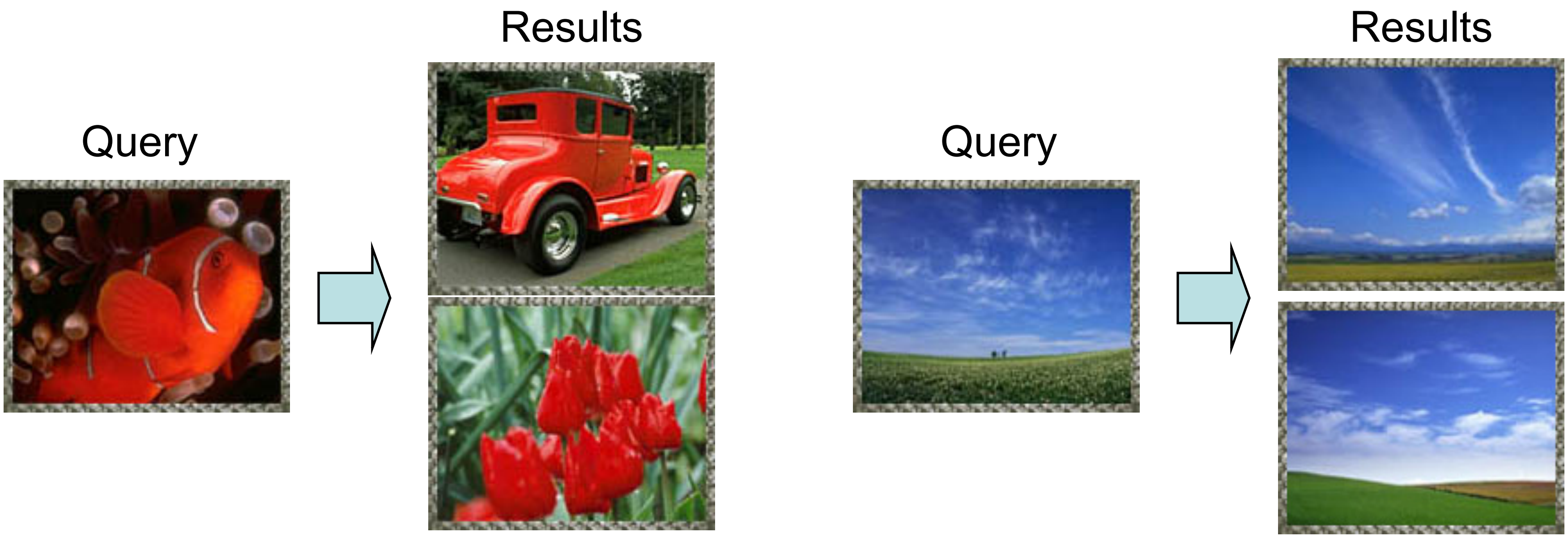
- Color histograms miss important spatial information
- Approaches
  - Apply color histograms to regions
    - Dynamic regions
    - Static boundaries
  - Treat the spatial distribution as another feature space
  - Look for shape features

Color and Shape Queries



- Good reproduction of color histogram search, but fails to capture much of the spatial distribution
- Still a very hard problem

Search by Example



- Because all selection is feature-based, we can use an image as the query
- Query generates a feature vector, which is then checked for similarity in the rest of the database

ImageNet Architecture

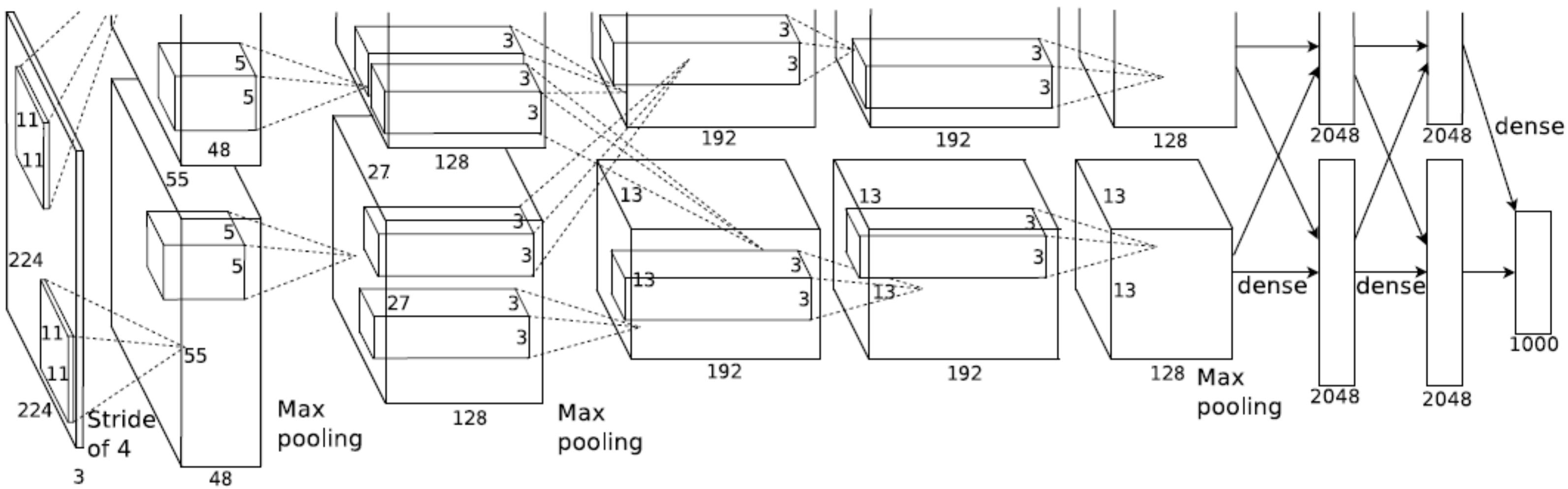


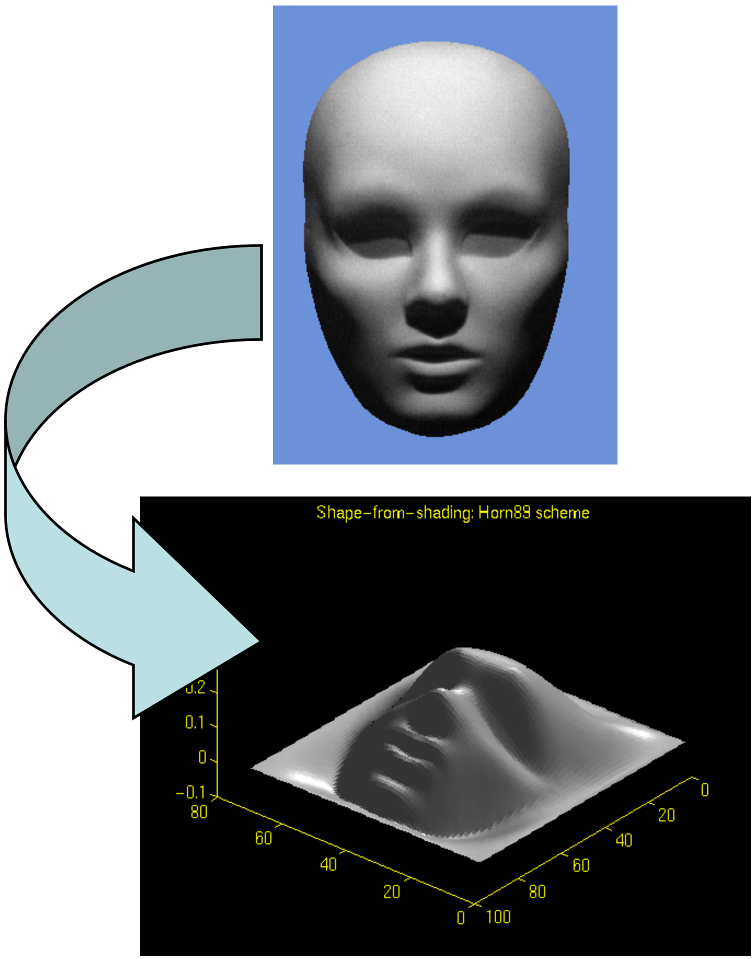
Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

Current (2018) best scores on ImageNet yield a 2% error rate.

Extracting 3D Structure: Shape From X

(where X= stereo, motion, focus, contours, texture, shading, etc.)

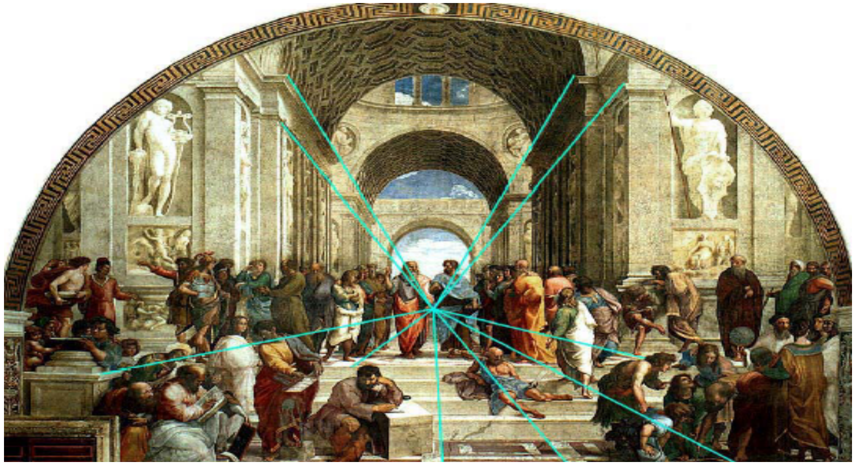
Shape from Shading



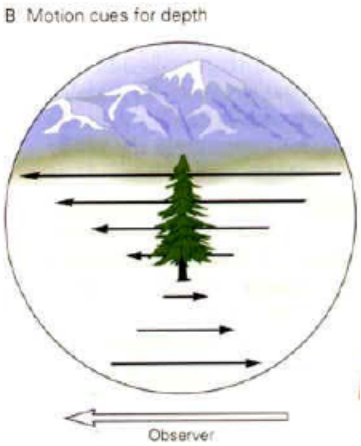
- Very hard!
- Requires a large amount of information and/or assumptions:
  - Smoothness assumption
  - Lighting and reflectance models
  - Structural information

Monocular Depth Cues

Texture gradient



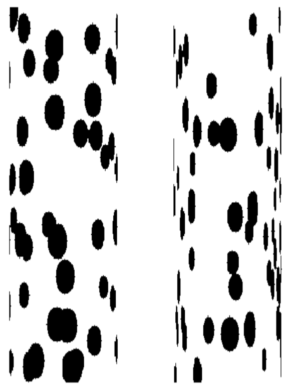
Linear perspective in Renaissance art (School of Athens, Raphael)

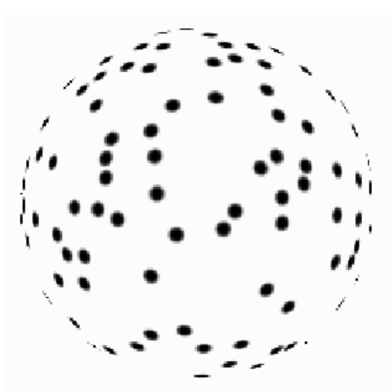


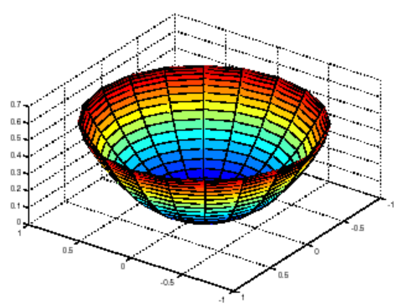
Motion parallax

Relative size/ familiar size

Shape from Texture

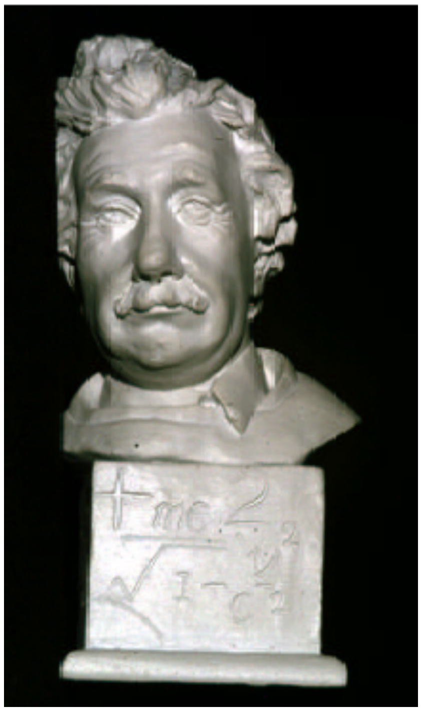


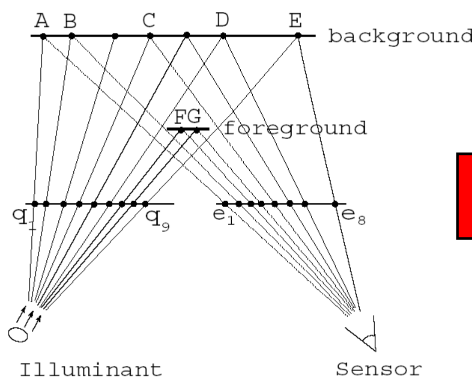




- Relies upon pre-defined knowledge of the texture pattern (since the repetition and distortion produces the shape)
- Mostly applied to synthetic images
- Very hard to specify in real-world images

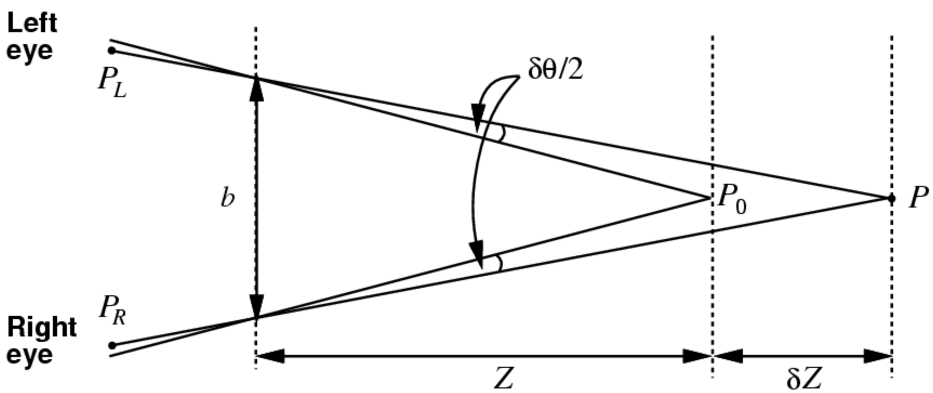
Shape from Structured Light (Creating a Known Texture)

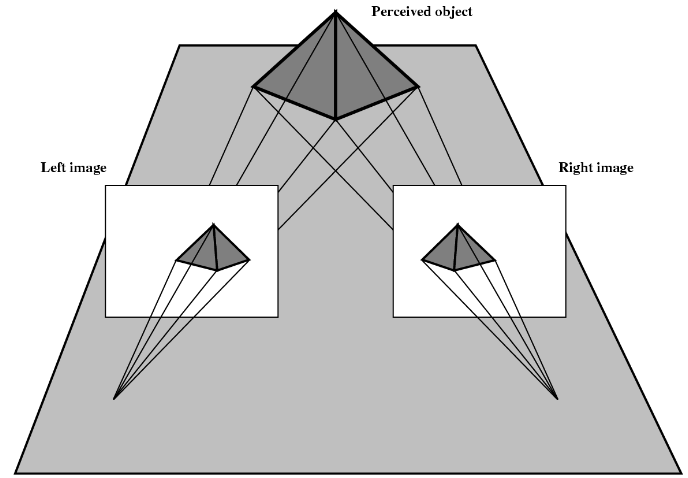




- Add structured light to a scene to make a known texture
- Look for the variations of this known texture to reconstruct 3D features

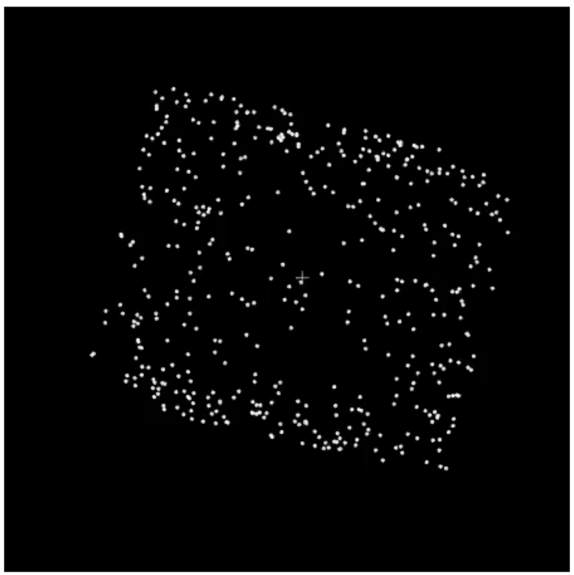
Relating Disparity and Depth

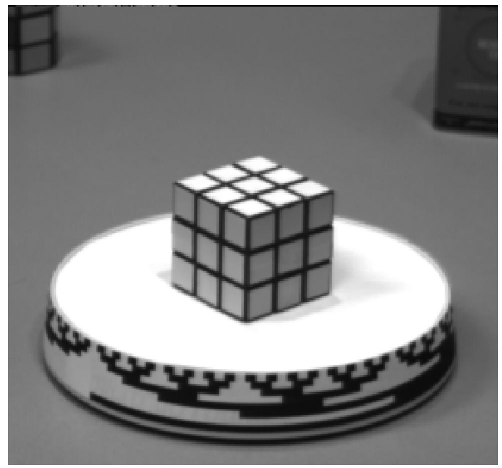


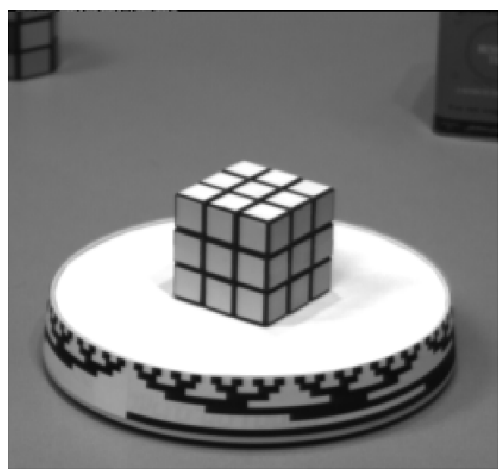


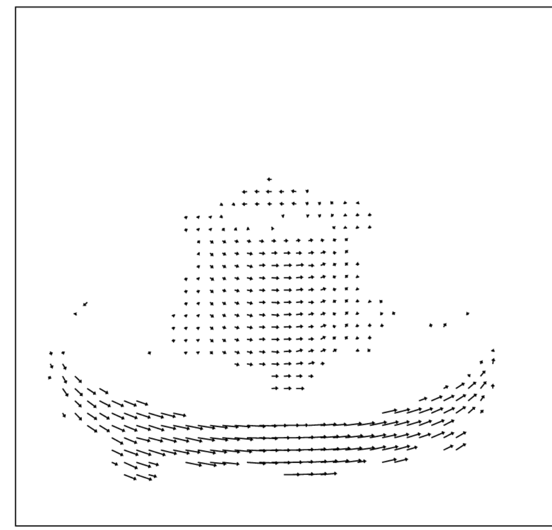
- Disparity is the difference between the locations of the two image projections
- If the camera positions are known then the interpretation relies upon matching features from the two images and doing the geometry

Shape from Motion

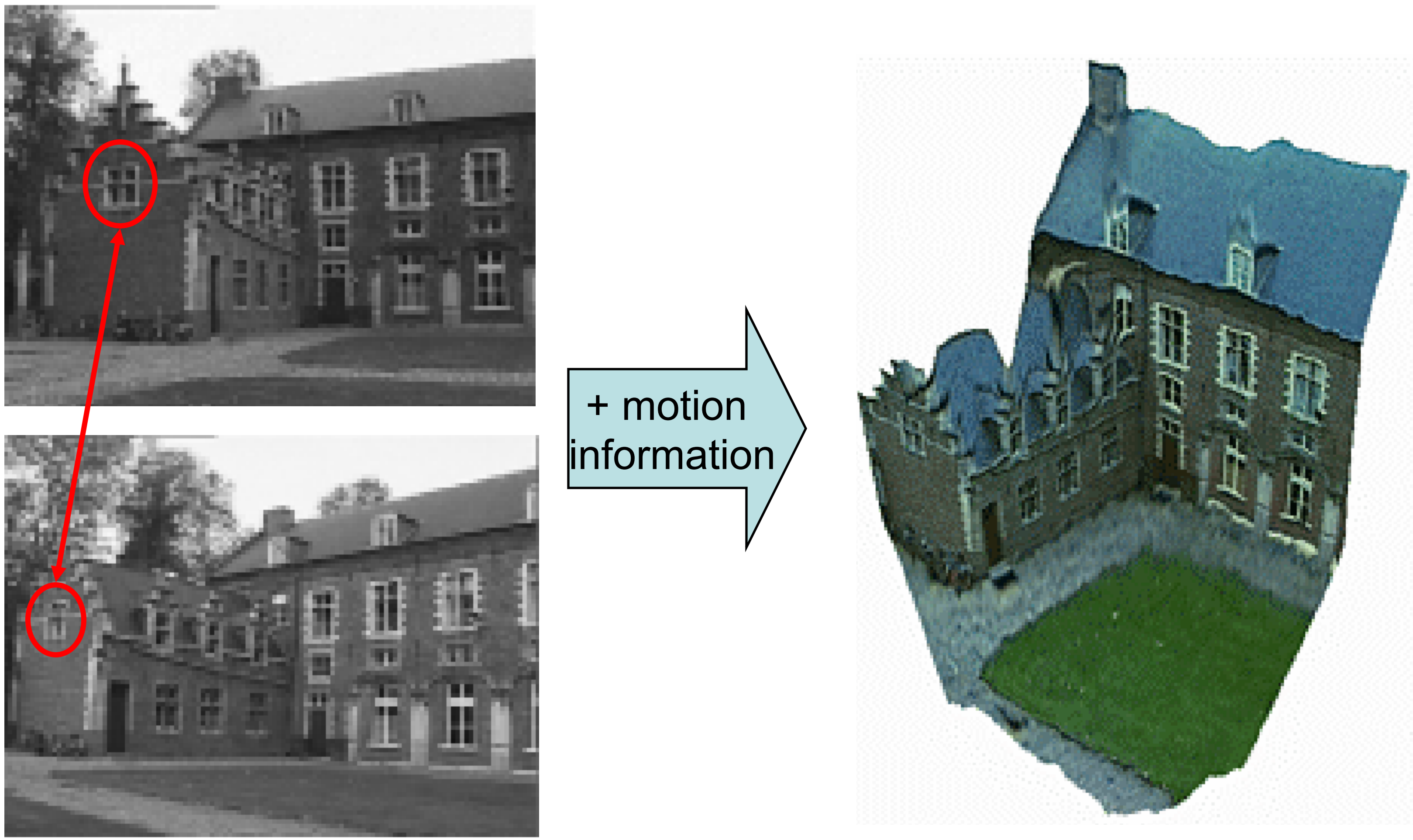




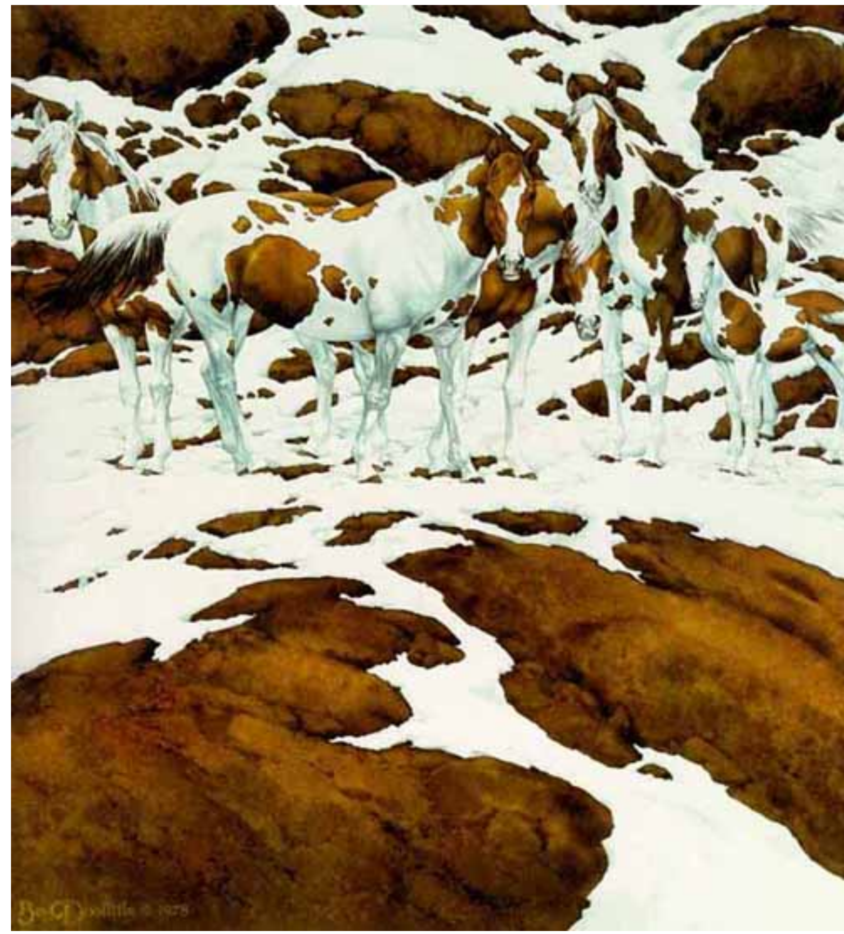




From Motion to 3-D Models



Context matters

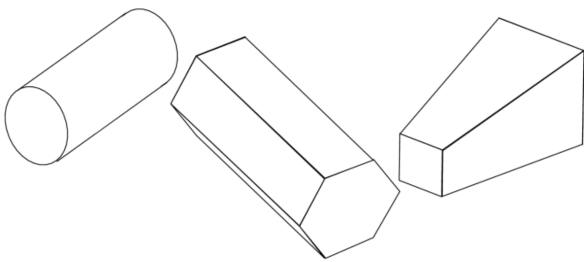


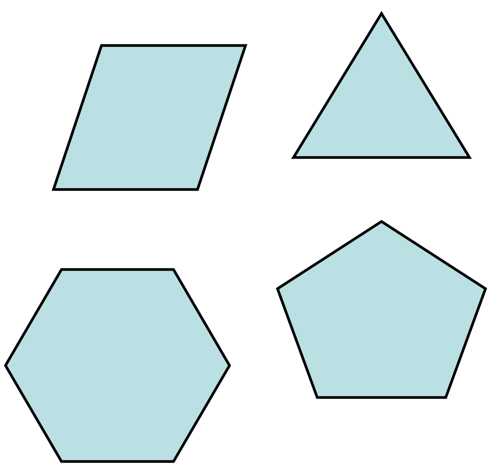
Object Recognition

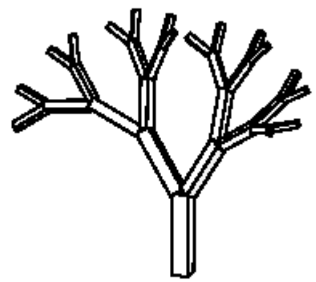
Two Schools of Thought

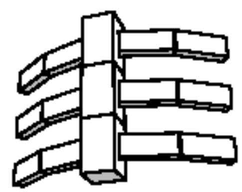
- Structural Pattern Recognition
  - The data is converted to a discrete structure (such as a grammar or a graph) and the techniques are related to computer science subjects (such as parsing and graph matching).
- Statistical Pattern Recognition
  - The data is reduced to vectors of numbers and statistical techniques are used for the tasks to be performed.

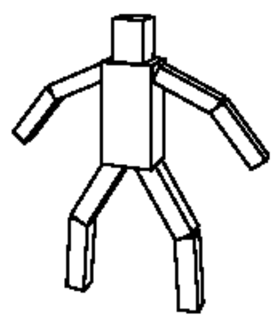
Structural Pattern Recognition







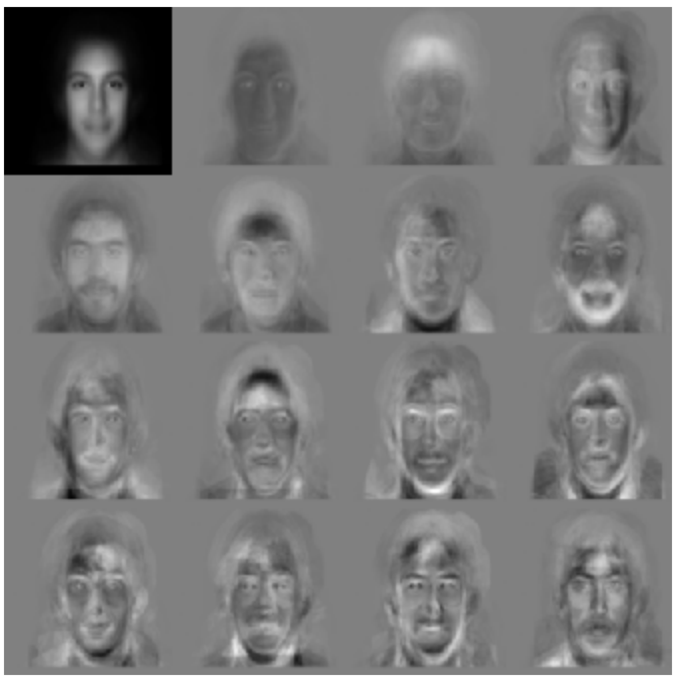




- (related to model-based methods)
- Generalized Object Representation
  - Generalized Cylinders
  - Polyhedral Representations
- Compose objects as a combination of these items
- What are the basic units?

Statistical Pattern Recognition

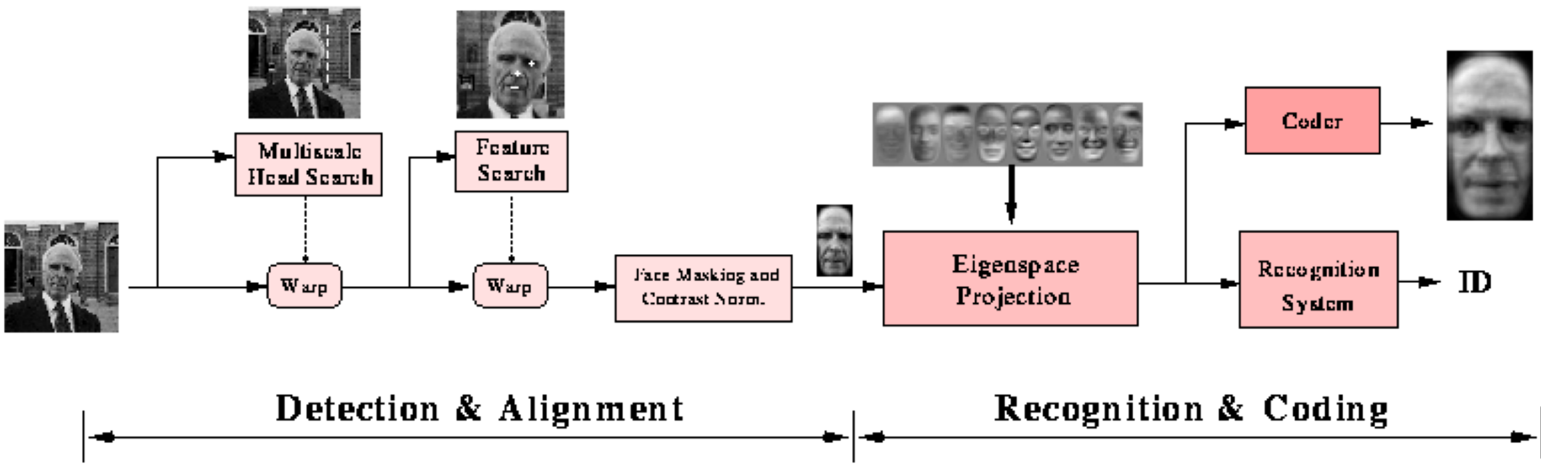
Standard Eigenfaces



Eigenfaces Projection

- Decompose the selected objects into a set of features
- Use statistical techniques to group the features into classes
- Assign detection and/or recognition probabilities based on statistical distribution
- Example from Takeo Kanade’s face detector

Face Detection and Recognition

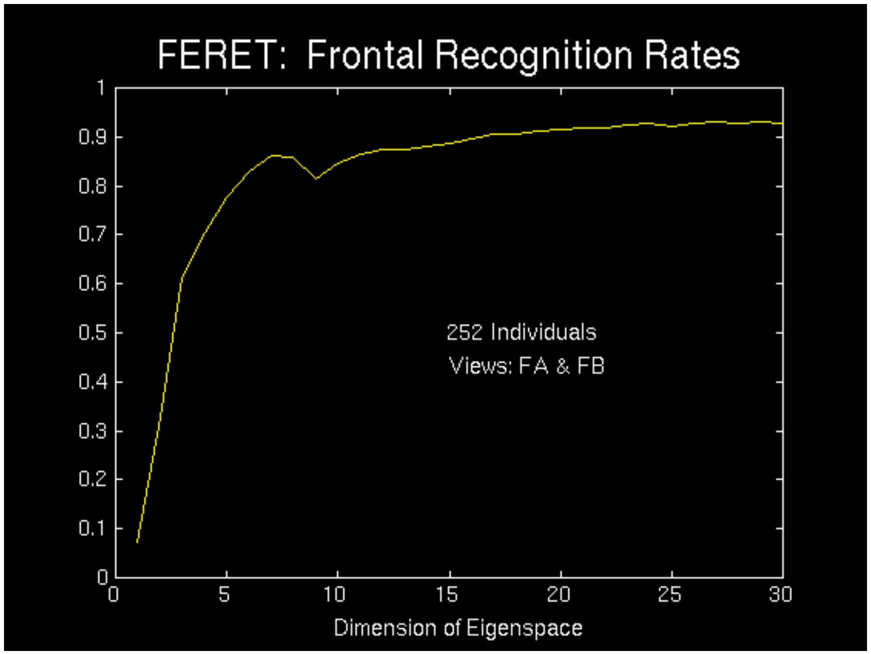


- Detection
  - Identify areas in the scene that are potential faces
- Alignment
  - Standardize the scale, rotation, and balance
- Recognition
  - Compare feature vector with stored databases

Face Detection

- Example from Takeo Kanade's face detector

Face Recognition (Pentland, MIT)



FERET face database

- A recognition accuracy of 99.35% was obtained using two frontal views of 155 individuals.

Deep Face

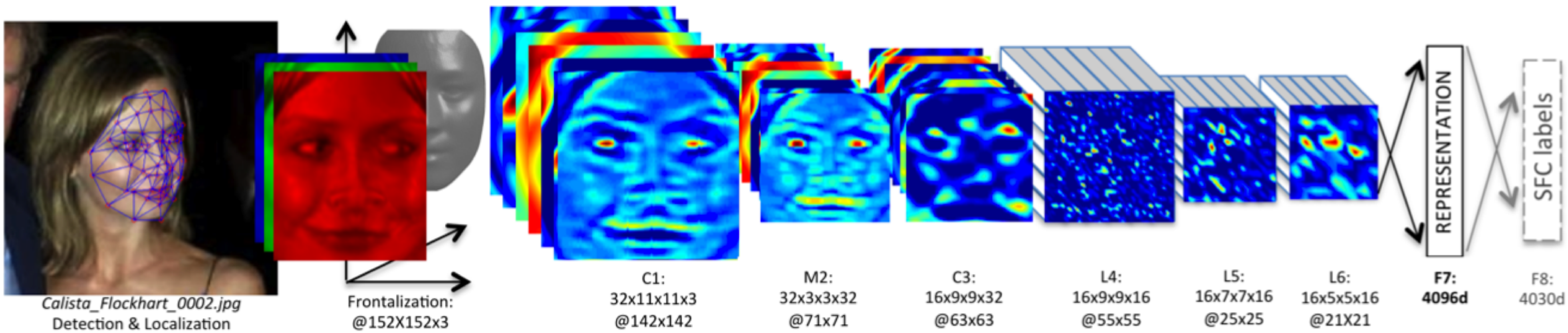


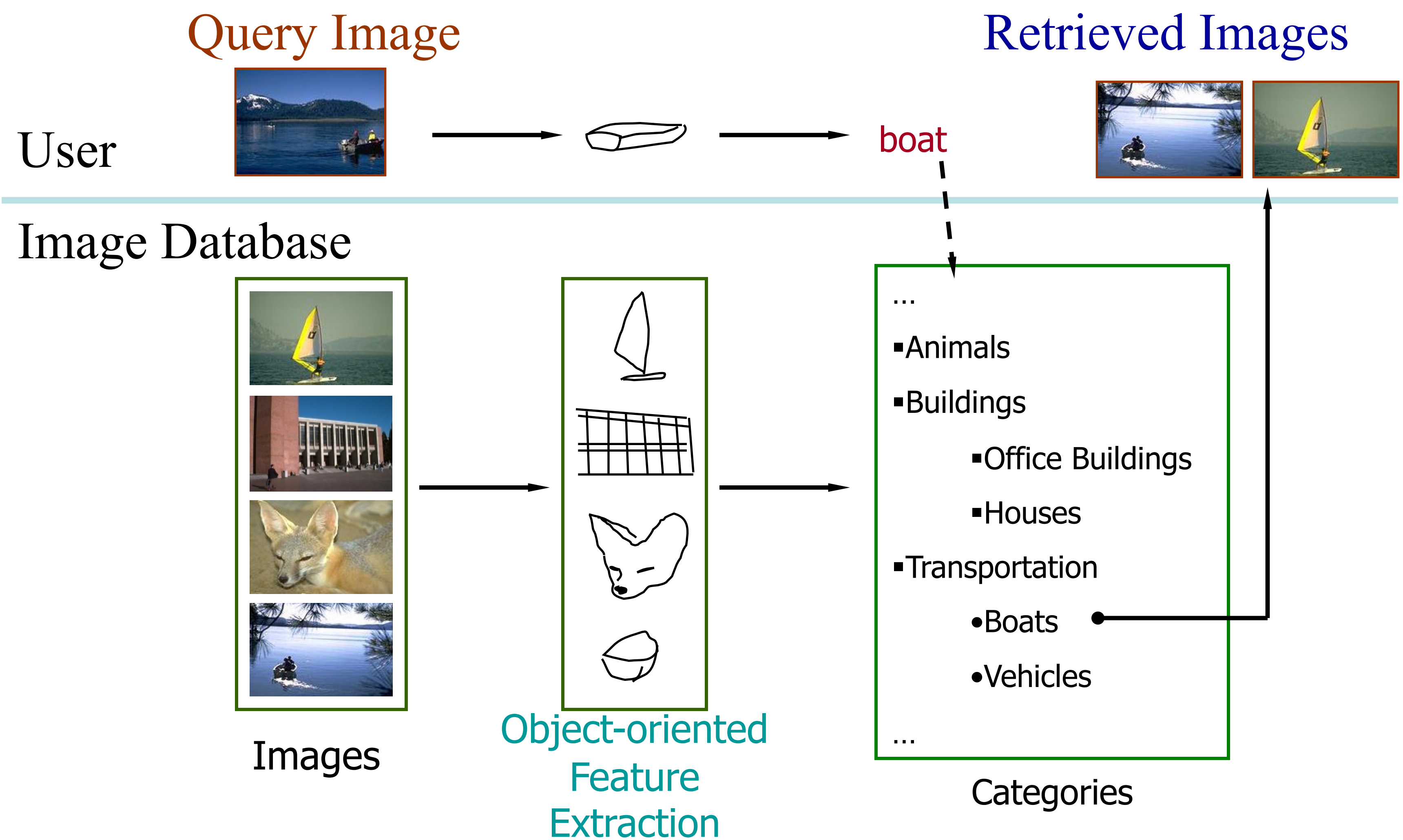
Figure 2. **Outline of the *DeepFace* architecture.** A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate feature maps produced at each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

Face Recognition: Detect → Align → Represent → Classify

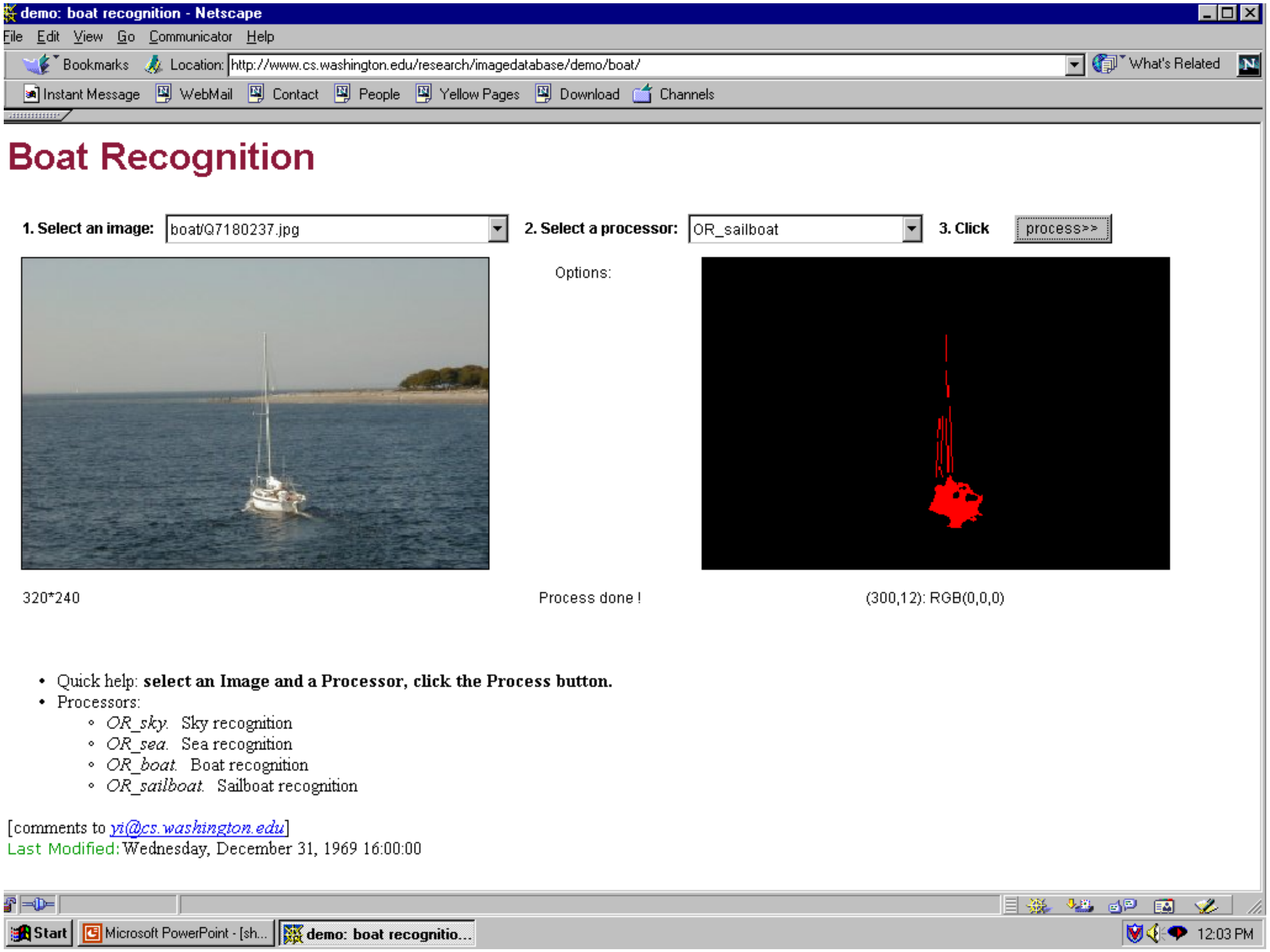
- Labeled Faces in the Wild dataset with 4M images
- Deep Face: 97.35% accuracy with 120M parameters
- Human performance: 97.5% accuracy

Taigman,Yang, Ranzato, & Wolf (Facebook), CVPR 2014

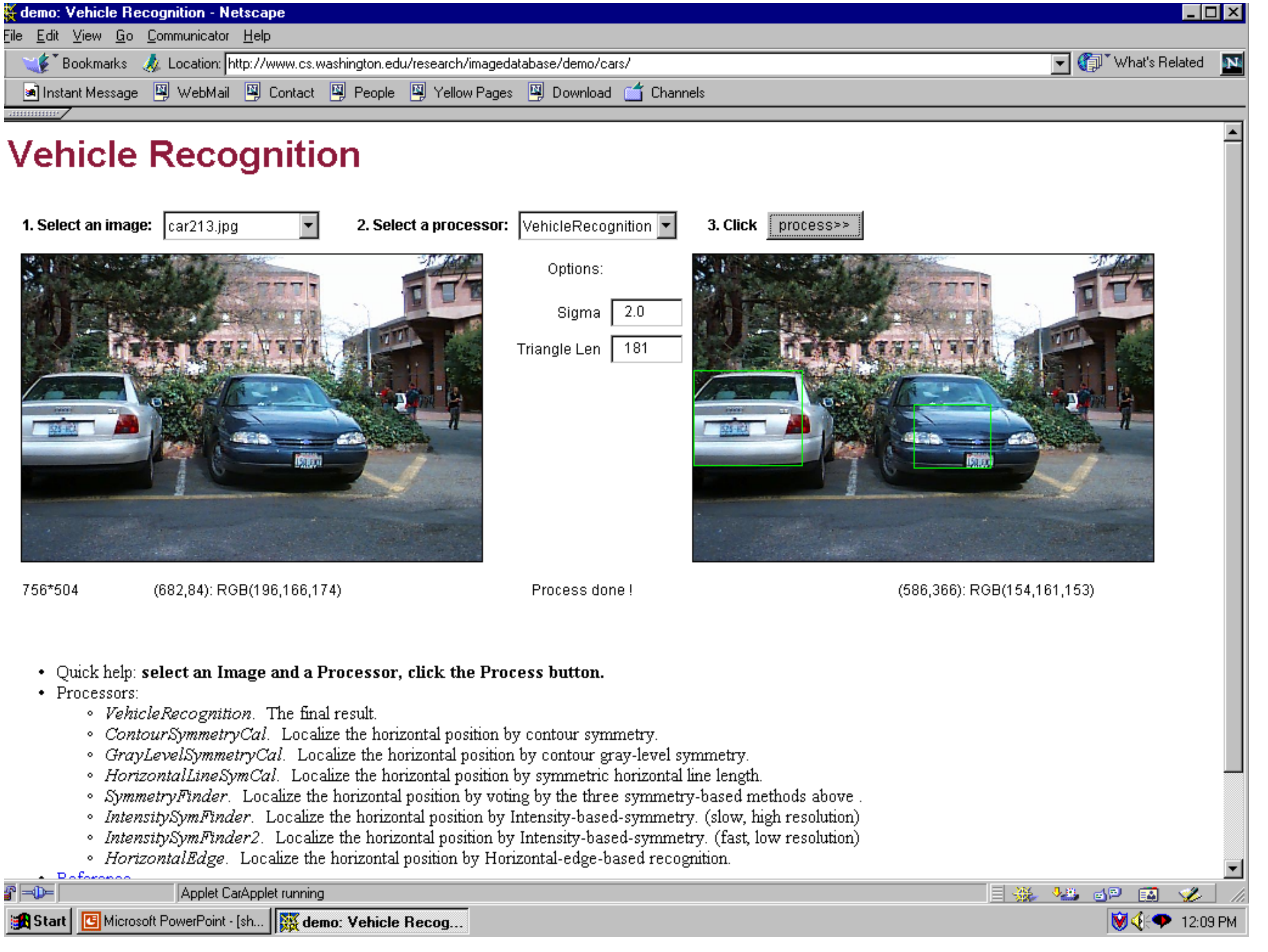
Detecting Arbitrary Objects Statistically: Building many, many Classifiers



Boat Detection



Vehicle Detection



Building Detection

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

Deep Visual-Semantic Alignments for Generating Image Descriptions (CVPSR 2015). Andrej Karpathy, Li Fei-Fei.

Today: Other Vision Tasks

- Finding similar images
- Extracting 3D structure
- Object recognition

Deep Network Problems