Experimental Design for Machine Learning on Multimedia Data Lecture 9

Prof. Gerald Friedland,
fractor@eecs.berkeley.edu

Website: http://www.icsi.berkeley.edu/~fractor/fall2019/

Today

- Some comments on projects: Generalization!

- Some old school but still useful knowledge on machine learning for audio, image and video

Next week!

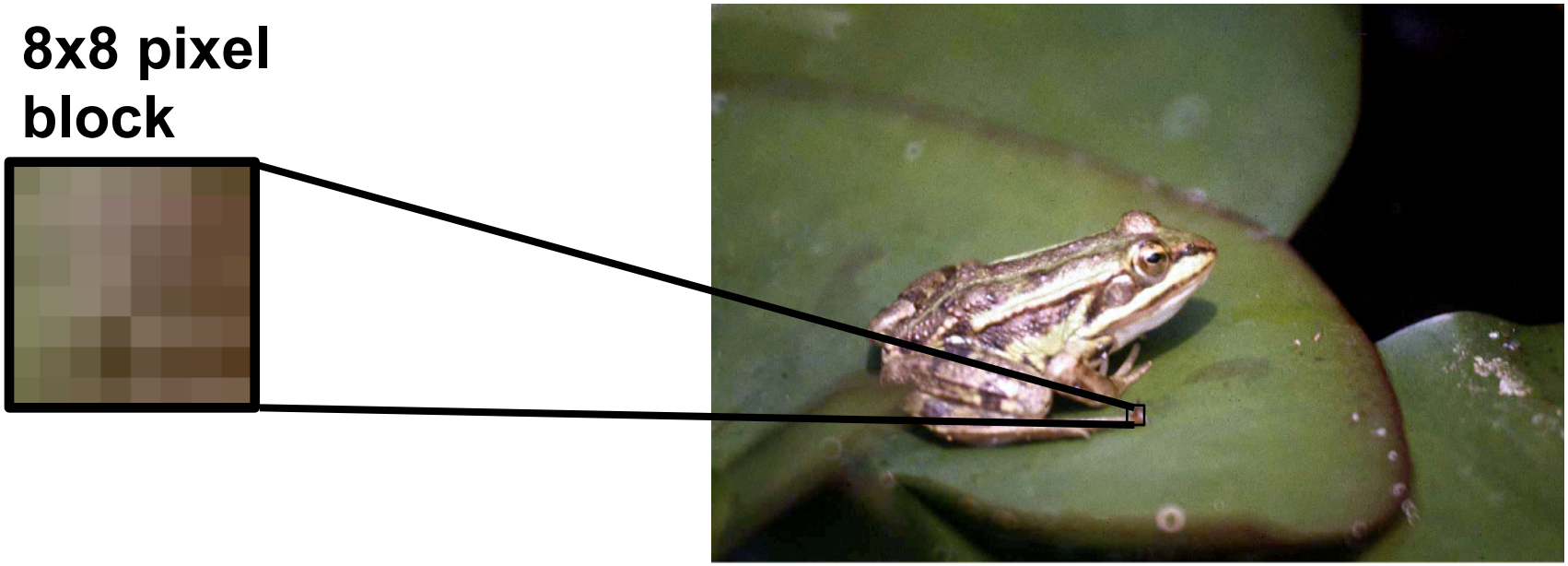
10-min presentations for each project

Please show your idea and your preliminary results in class!

What's the Problem?

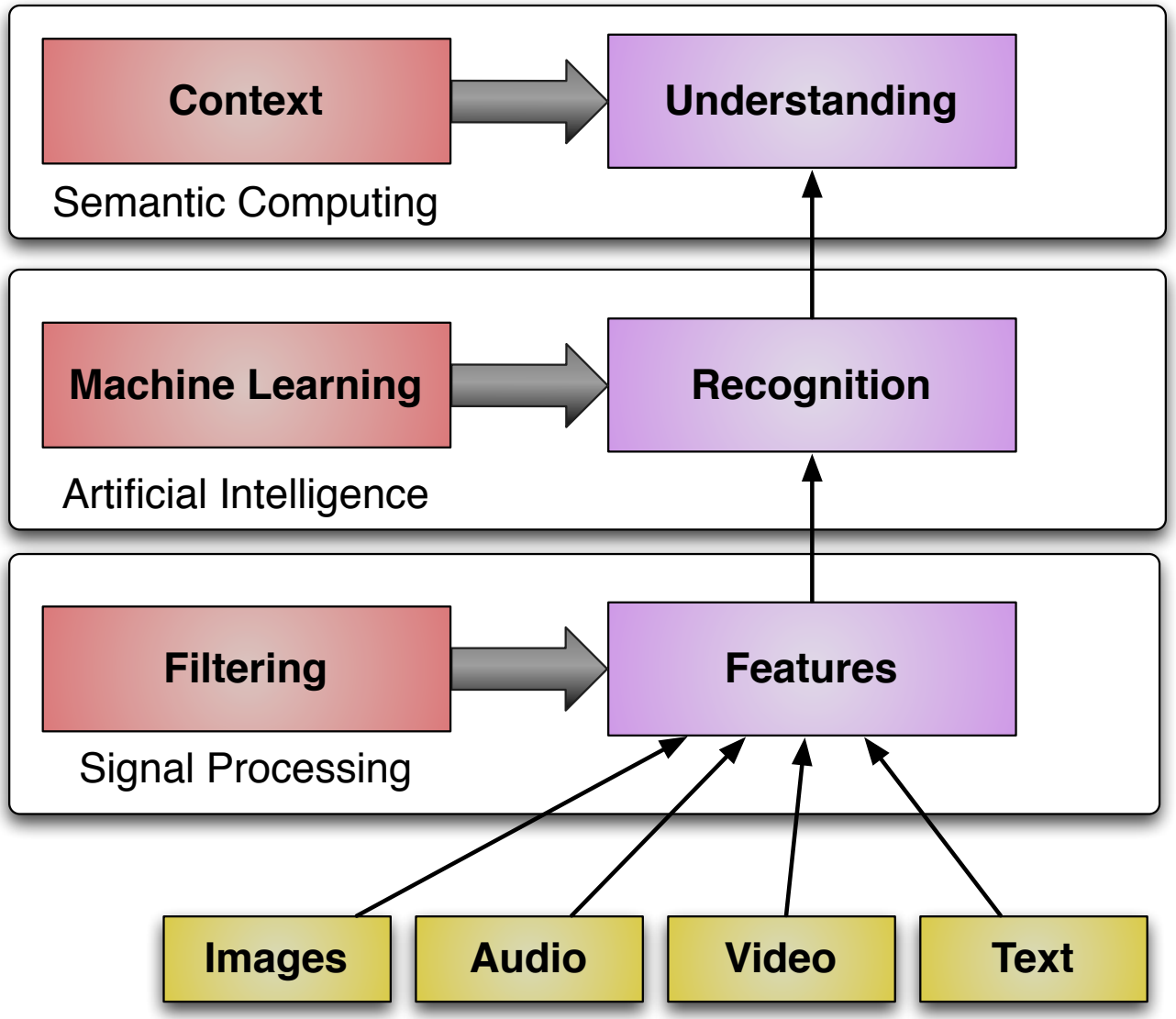
An Image:

8x8 pixel block



Source: bigfoto.com

Multimedia Analytics



Basic Terms

- Digital image data

  A **pixel** (PICture ELement) is a sample of the image intensity quantized to an integer value. An image is a two-dimensional array of pixels.

- Color space

  Perceptually non-uniform: **RGB**, CMYK

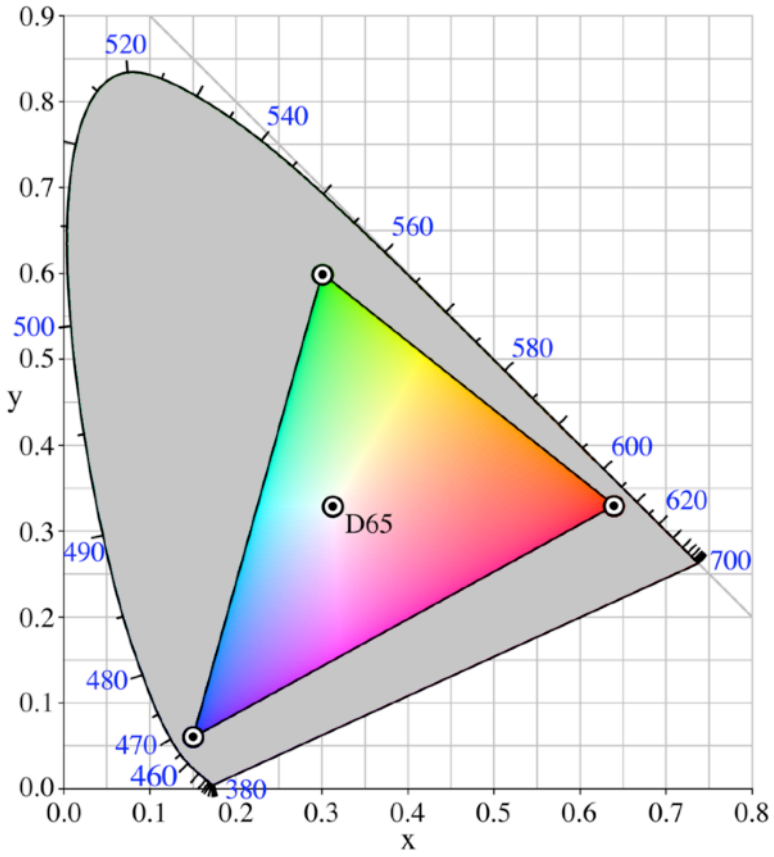
  Perceptually “uniform”: HSV, CIELab, CIELuv, etc.

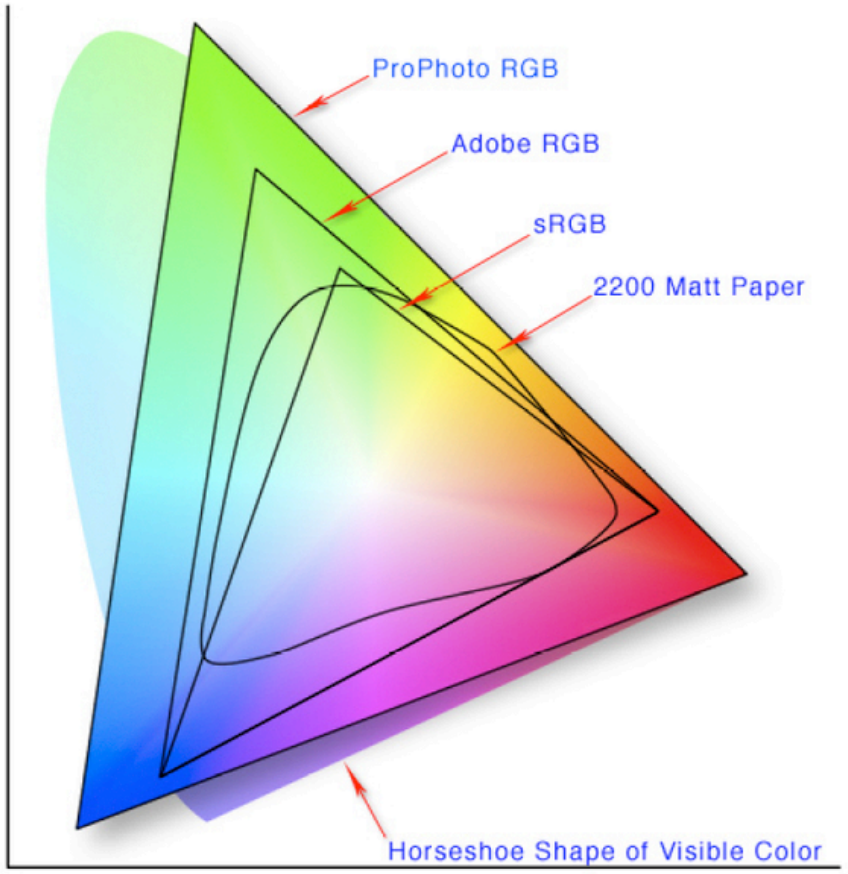
RGB Space

- RGB (Red-Green-Blue) encoding in graphic systems usually uses three bytes enabling $(2^8)^3=2^{24}$ or roughly 16 million color codes

- Each 3-byte or 24-bit RGB pixel includes one byte for each of red, green, and blue e.g., red (255,0,0), green (0,255,0), blue (0,0,255), black(0,0,0), white(255, 255, 255), etc.

RGB Color Space



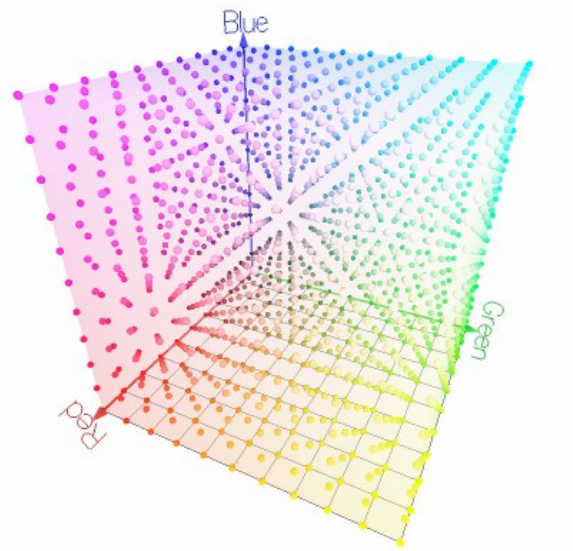


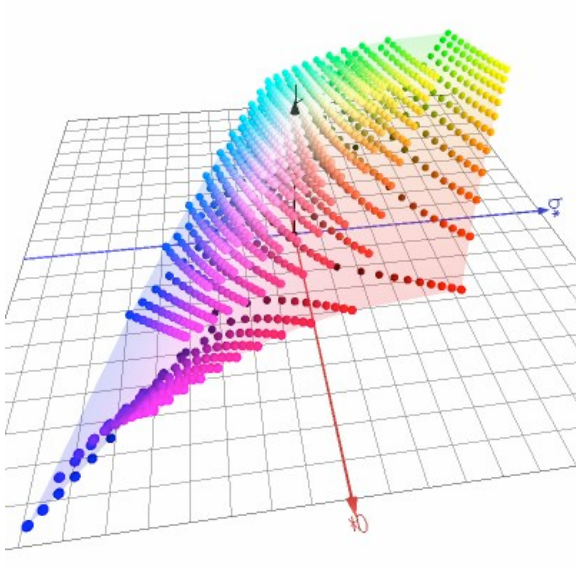
- RGB (Red-Green-Blue) encoding three bytes enabling $(2^8)^3=2^{24}$ colors.

Source: Wikipedia

Other Color Spaces

- CMY(K): cyan, magenta, yellow, (and black)
- YCbCr
- HSV (hue, saturation, and value)
- CIE L*a*b*, CIE L*u*v*



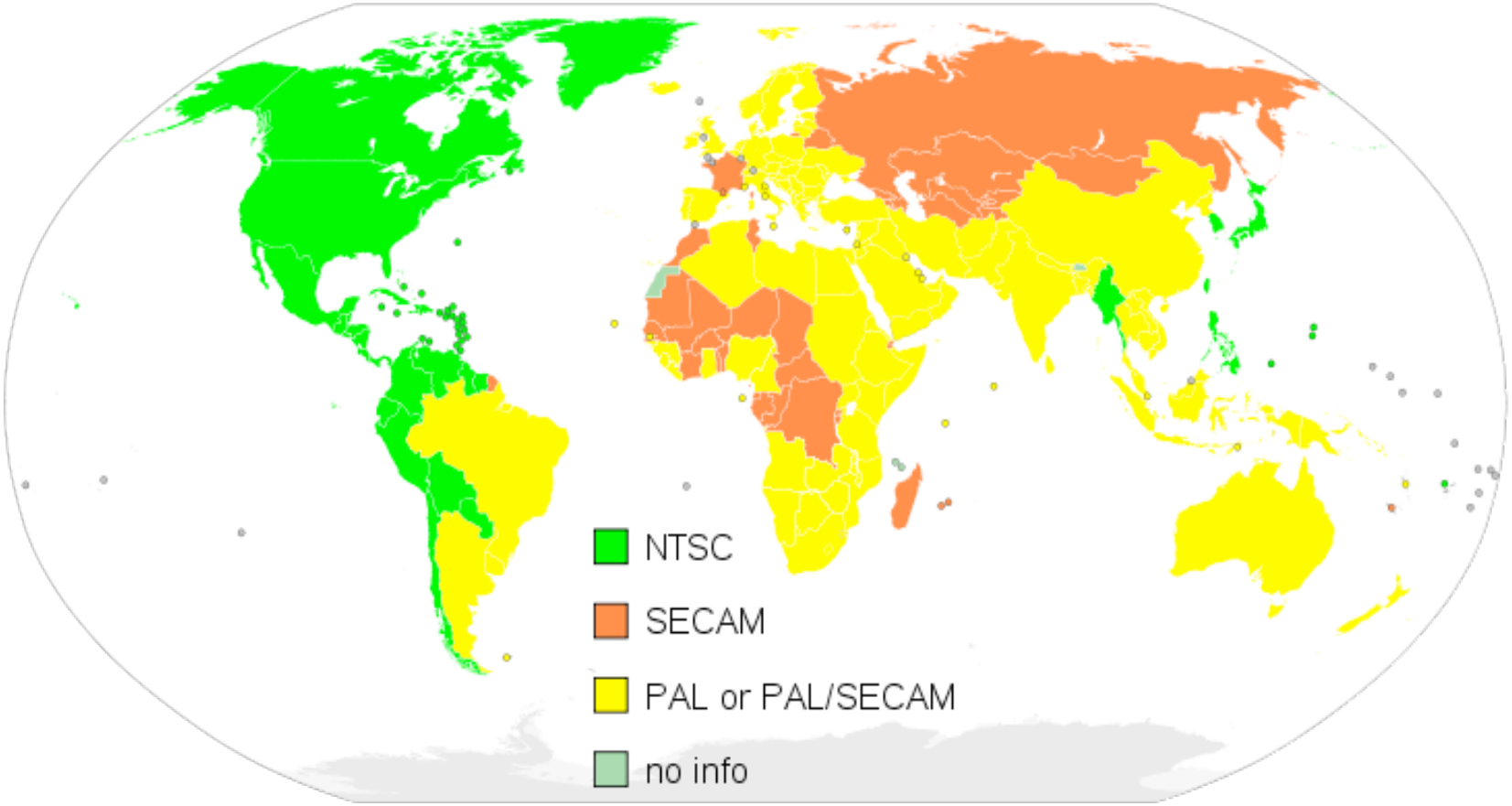


More Basic Terms

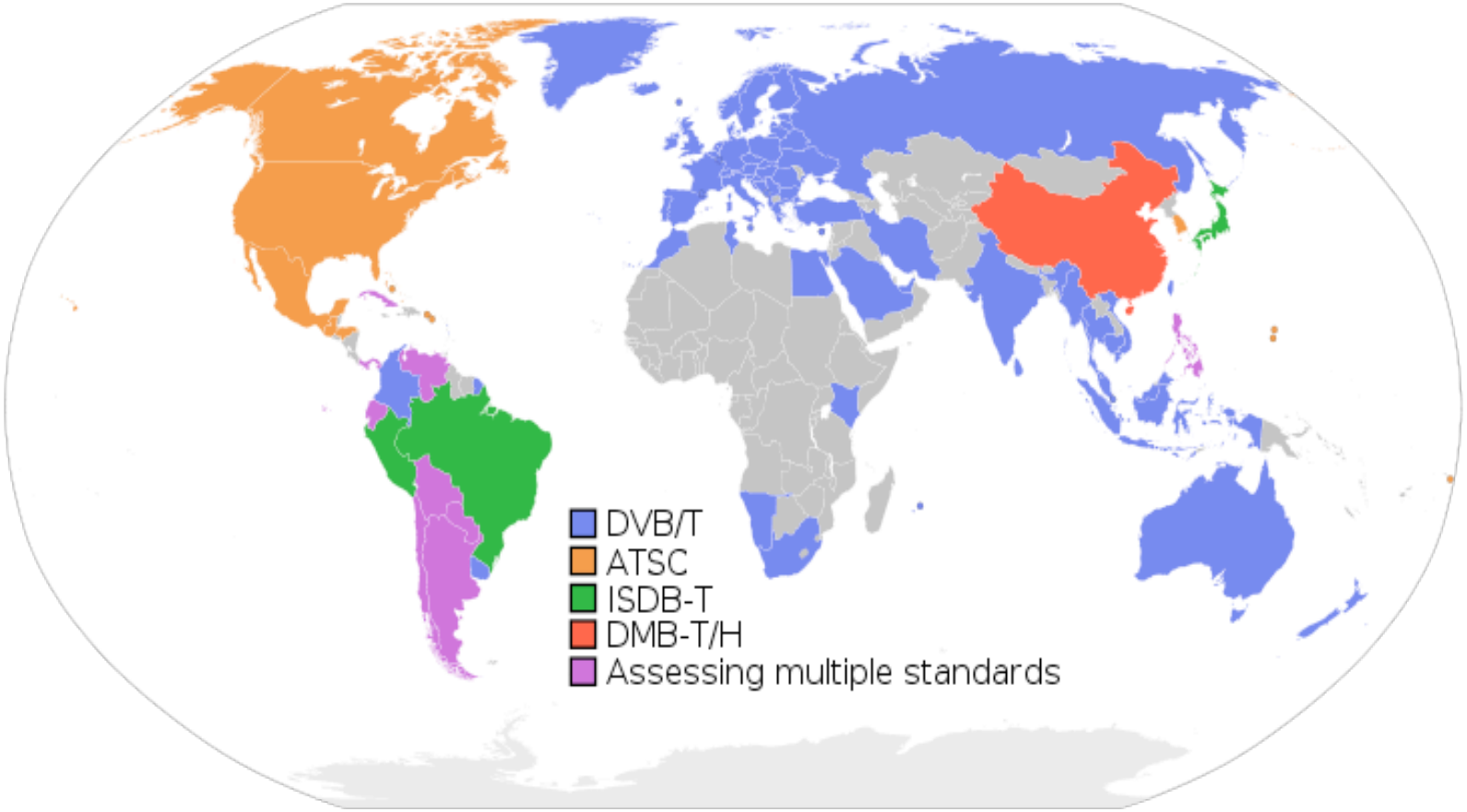
- Video Recording:
  The technology of electronically capturing a sequence of still images (frames) to represent scenes in motion

- Technologies:
  - Video cameras / analog cameras
    e.g., PAL (China), NTCS (USA), etc.
  - Digital cameras
    e.g., DMB (China), ATSC (USA), etc.

Global Distribution of Formats (analog)



Global Distribution of Formats (Digital)



Video Storage

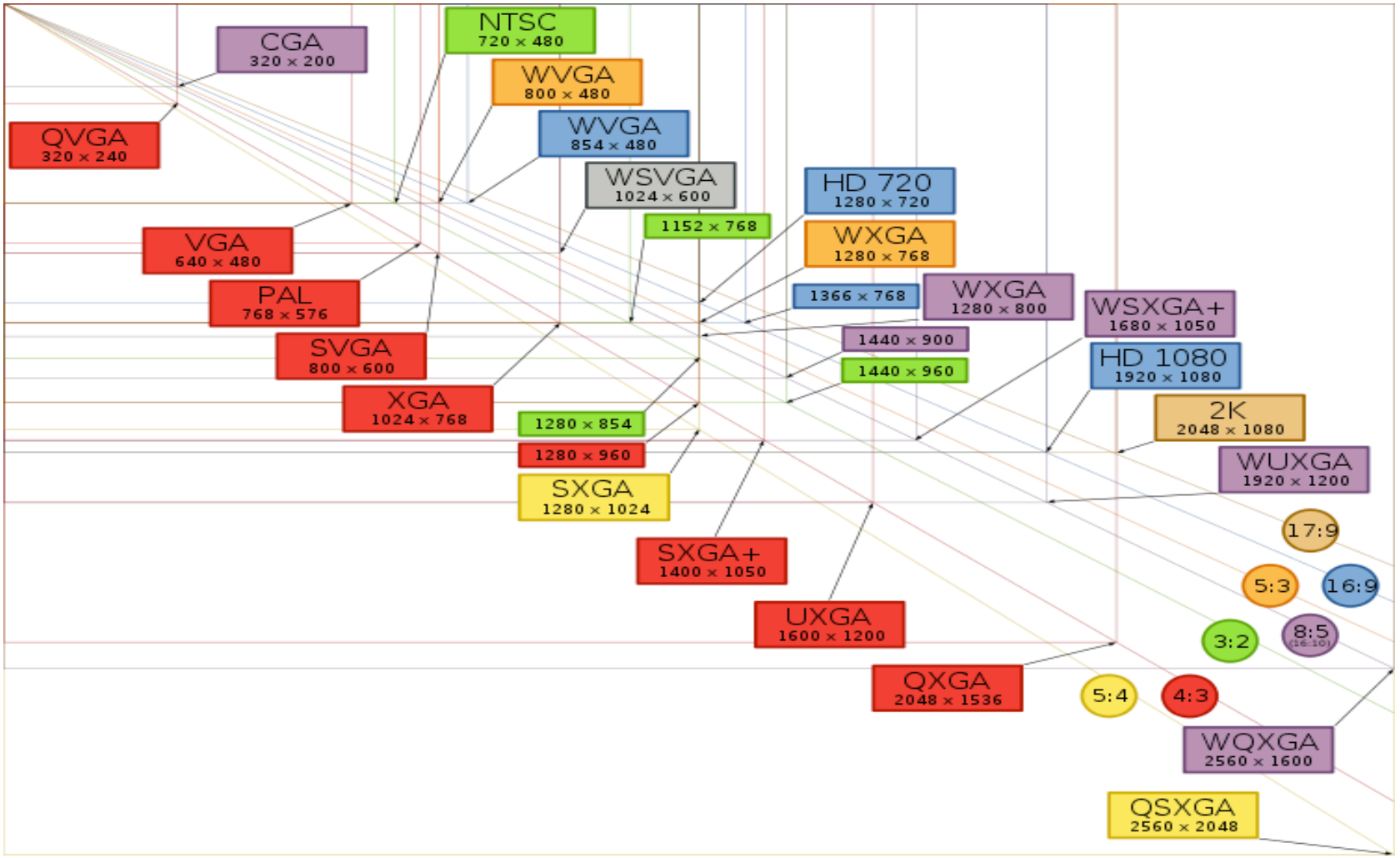
- Analog tape, e.g. VCR, Betamax (Sony), VHS (JVC), etc.
- Digital tape, e.g. DV, HDV, etc.
- Optical disc storage, e.g. VCD, Blu-ray Disc (Sony), DVD (Super Density Disc), etc.

More Basics

- Frame rate:
  the number of still pictures per second e.g.,
  25 frames/s (China), 29.97 frames/s (USA)

- Resolution:
  the number of distinct pixels in each dimension that can be displayed. Aspect ratio of an image is its width divided by its height.

Resolution and Aspect Ratio



Digital Video Formats

- Raw video / uncompressed video:

  e.g. Sony D-1 (1986), Apple QuickTime (1990), etc.

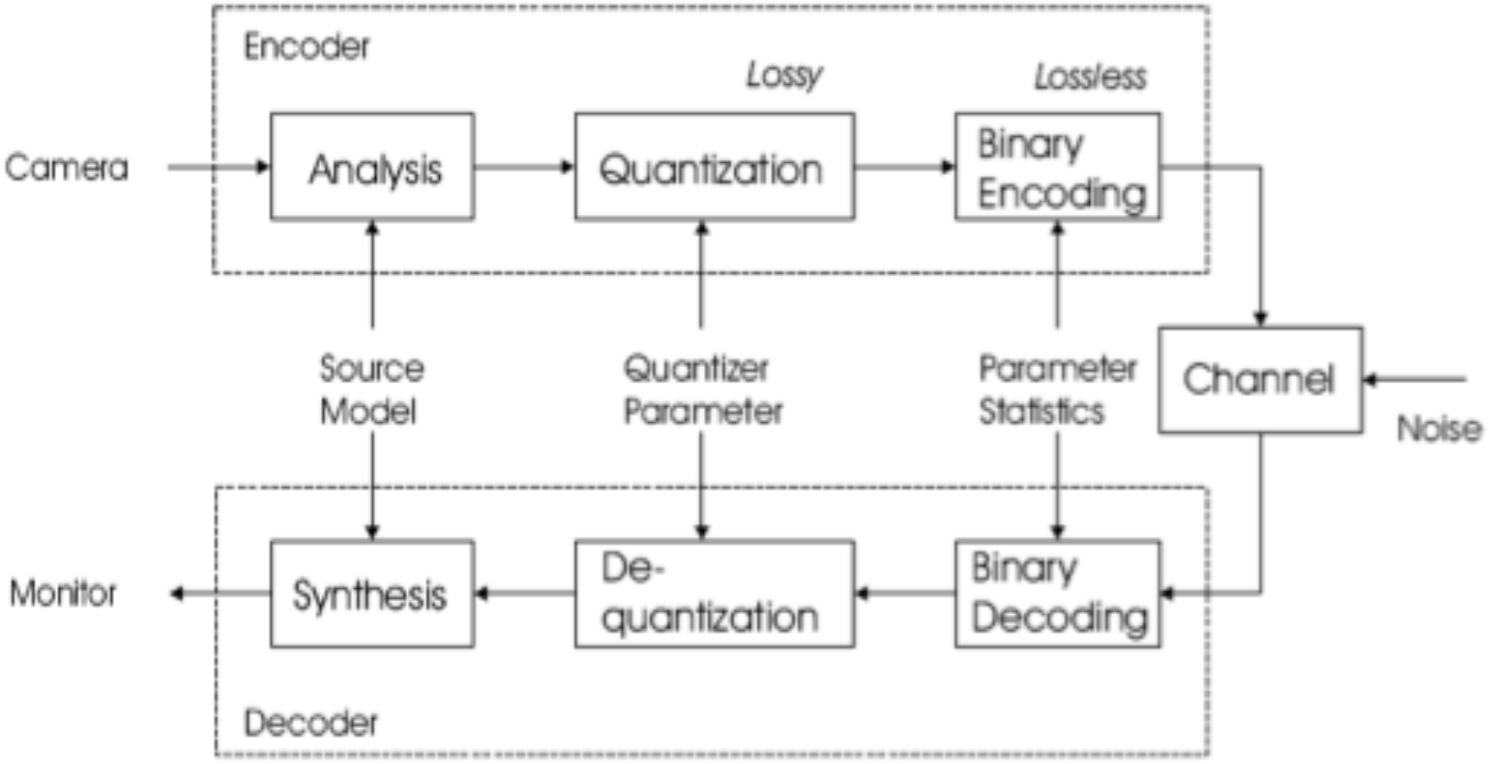
- Compressed video:

  e.g. MPEG-1 (1990), MPEG-4 (1998), MPEG-7 (2000). Most common now: H.264.

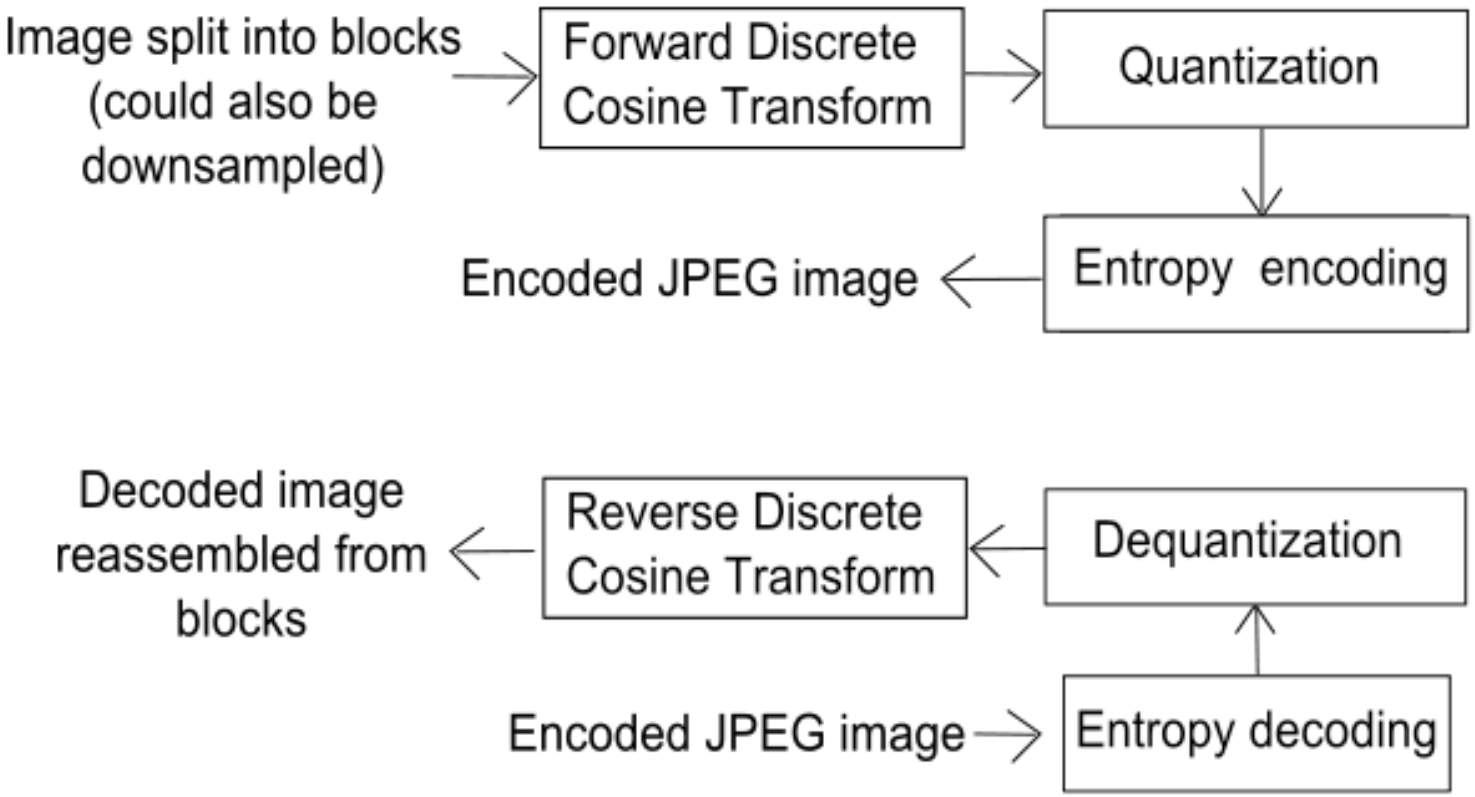
- 3-D video:

  digital video in three dimensions (MPEG-4 Part 16 Animation Framework eXtension

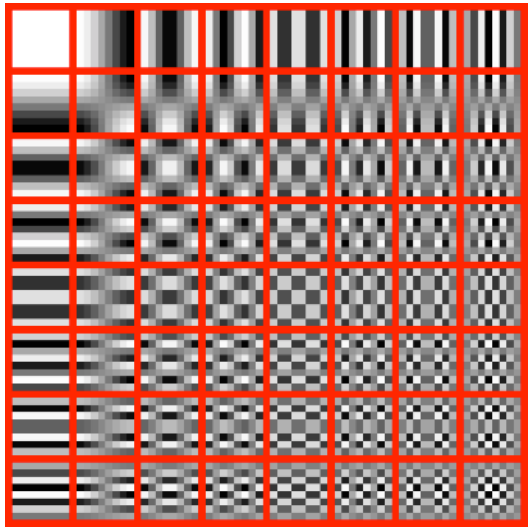
Video Coding



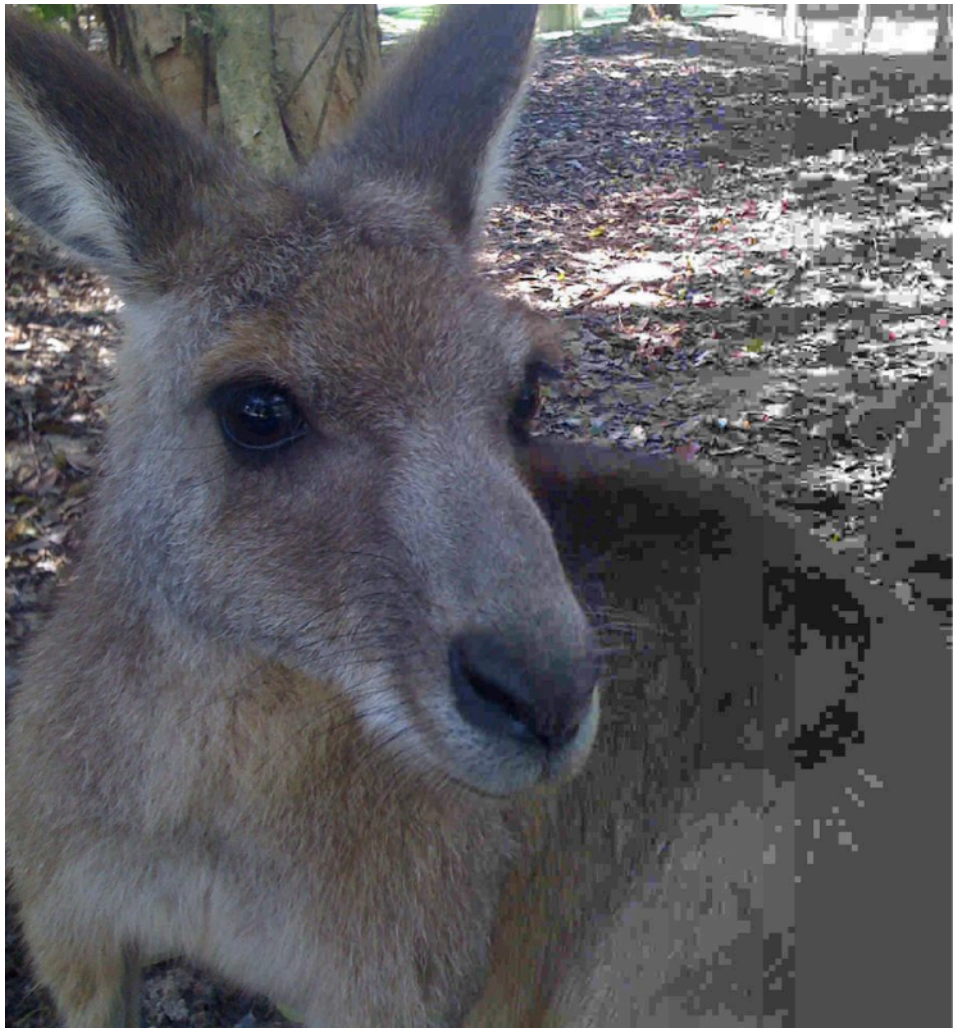
Encoding (Frame Coding)



Encoding (Frame Coding)



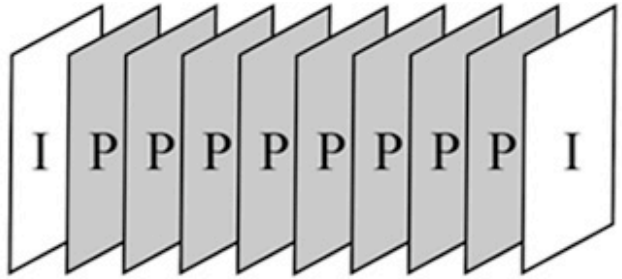
$$\begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$



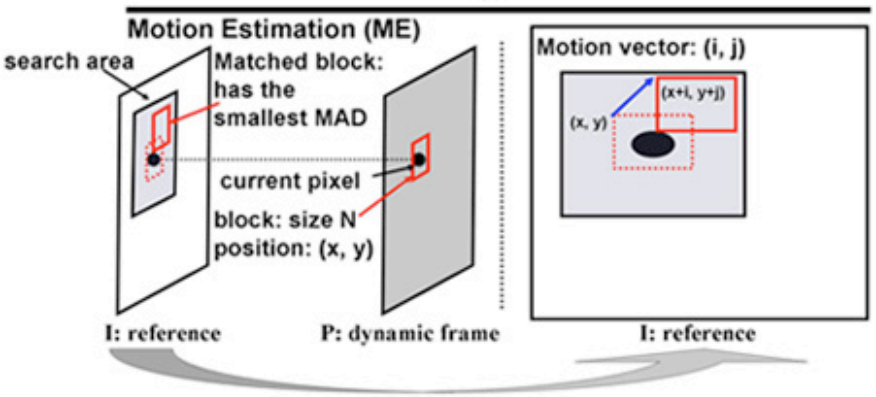
Encoding (Motion Compensation)

- Frame type (compressed video)
  - I frame: ‘intra-coded picture', in effect a fully-specified picture, like a conventional static JPEG file
  - P frame: ‘predicted picture‘, holds only the changes in the image from the previous frame
  - B frame: ‘bi-predictive picture‘, saves even more space by using differences between the current frame and both the preceding and following frames to specify its content

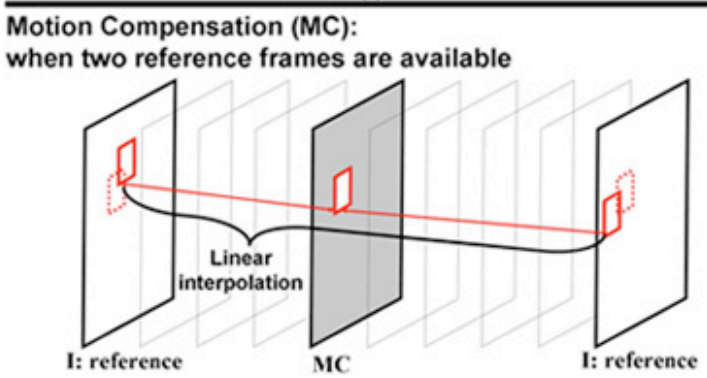
From Image to Video



(a)

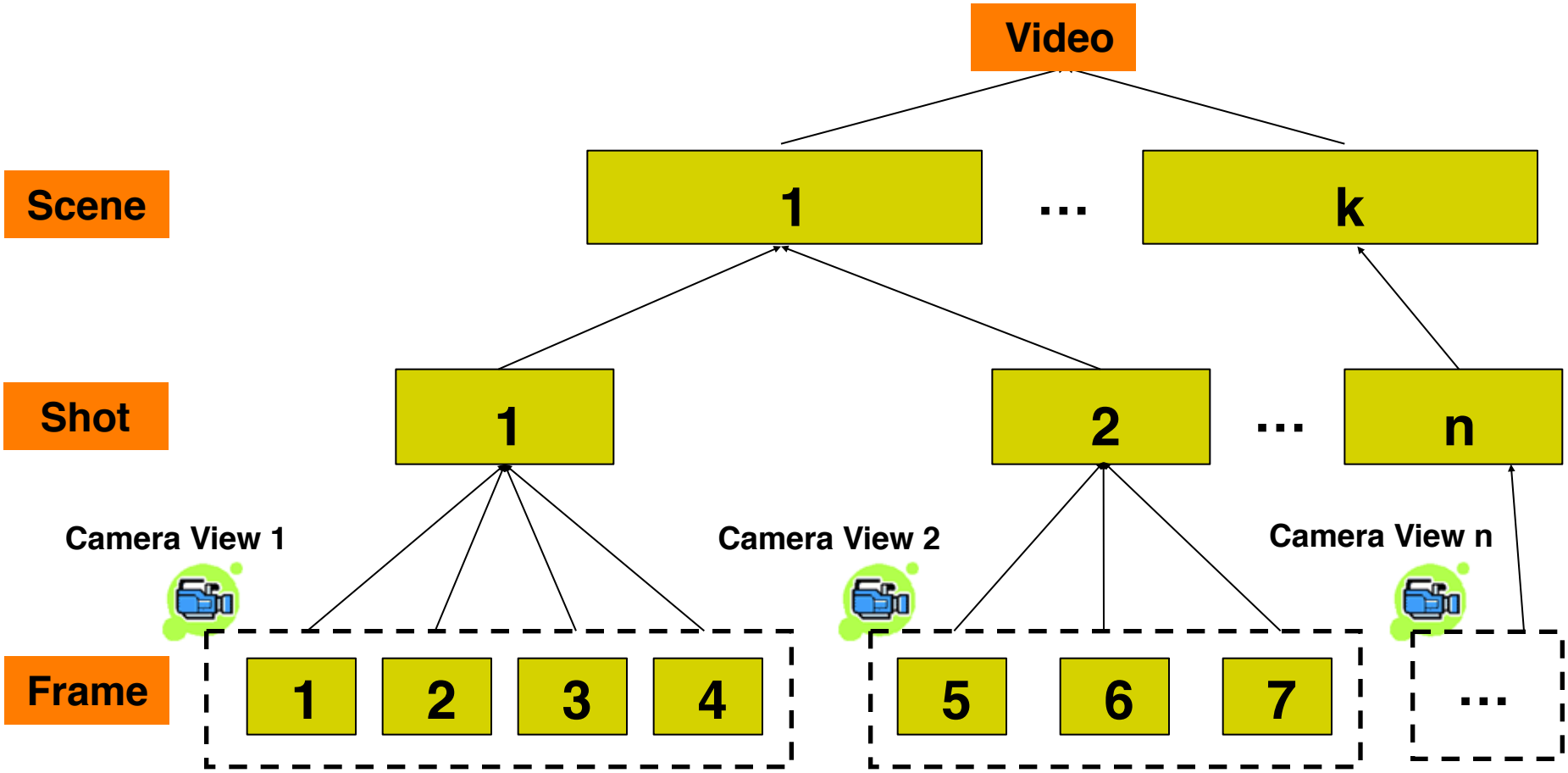


(b)



(c)

Content Structure of Video



Typical Video Features

Low-level features:

- Color features: color dominant, color histogram, color moment, etc.
- Texture features: structural features, statistical features
- Shape features: edge detectors, boundary-based, region-based, etc.

Remember: Audio Features

- Low-level features
  - Time-domain features:
    Volume, ZCR (Zero Crossing Rate), etc.
  - Frequency-domain features:
    Pitch, FC (frequency centroid), Energy, F0 (fundamental frequency), MFCC (Mel-frequency ceptral coefficients), etc.

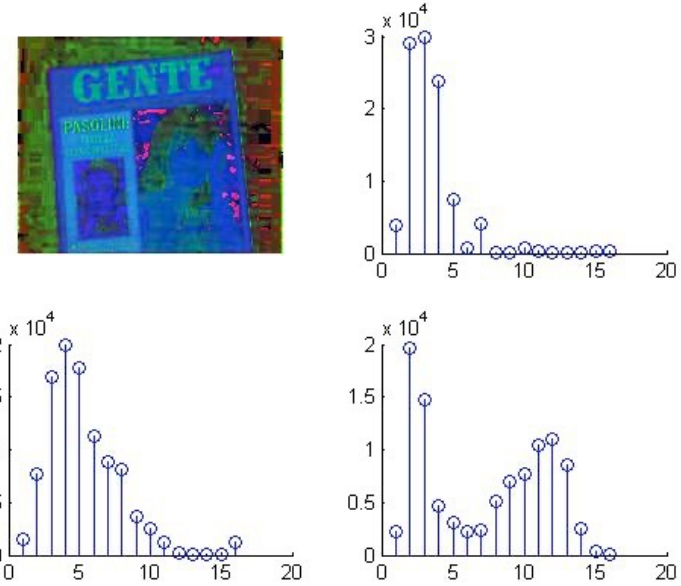
Audio Features

- Middle-level features:
  - Categories of sound: silence, music, environment, and speech
  - Different music types: songs, rock, piano, jazz, classical music, pop music, etc.
  - Various environment sounds: crowds, laughter, machine, telephone, etc.
  - Speech: male speech, female speech, diarization, speech recognition, etc.

Simple Visual Features

- Low-level features
  - Color features:
    dominant color, color histogram, color moment, etc.
  - Texture features:
    structural features, statistical features
  - Shape features:
    edge features, boundary-based, region-based

Color Histogram




```
hsv_i = rgb2hsv(my_image);
subplot(221), imshow(hsv_i);
[rHist,rcount] = imhist(hsv_i (:,:,1),16);
subplot(222), h1 = stem(1:16, rHist);
[gHist,gcount] = imhist(hsv_i (:,:,2),16);
subplot(223), h2 = stem(1:16, gHist);
[bHist,bcount] = imhist(hsv_i (:,:,3),16);
subplot(224), h3 = stem(1:16, bHist);
```

Simple Visual Features

- Color moments
  - First order: mean
  - Second order: variance
  - Third order: skewness
  - Forth order: kurtosis

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} f_{ij}$$

$$\sigma_i = \left(\frac{1}{N}\sum_{j=1}^{N}(f_{ij}-\mu_i)^2\right)^{\frac{1}{2}}$$

$$s_i = \left(\frac{1}{N}\sum_{j=1}^{N}(f_{ij}-\mu_i)^3\right)^{\frac{1}{3}}$$

$$k_i = \left(\frac{1}{N}\sum_{j=1}^{N}(f_{ij}-\mu_i)^4\right)^{\frac{1}{4}}$$

Visual Features

- Texture features
  - Texture gives information about the spatial arrangement of the colors or intensities in an image
  - Structural features: structural primitives and their placement rules
  - Statistical features: statistical distribution of the image intensity

Simple Visual Features

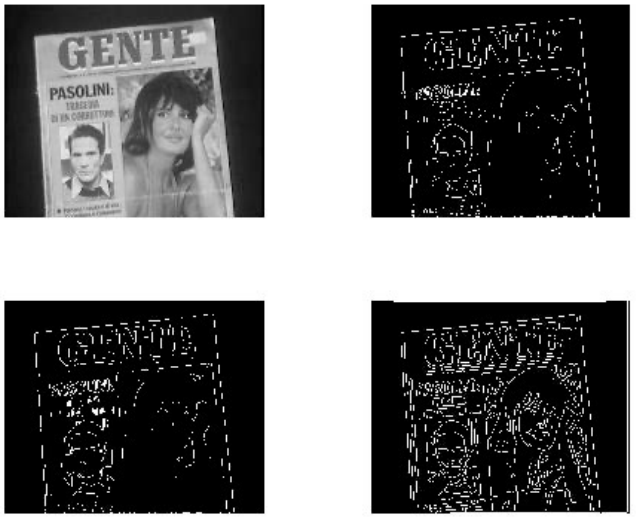
- Texture features
  - Gray level co-occurrence matrix (GLCM) $P_d[i,j] = n_{ij}$
  - Max probability $C_m = \max_{i,j} P_d[i,j]$
  - Energy $C(k,n) = \sum_i \sum_j P_d[i,j]^2$
  - Contrast $C(k,n) = \sum_i \sum_j (i-j)^k P_d[i,j]^n$
  - Entropy $C_e = -\sum_i \sum_j P_d[i,j] \ln P_d[i,j]$
  - Homogeneity $C_h = \sum_i \sum_j \frac{P_d[i,j]}{1+|i-j|}$

$$C_c = \frac{\sum_i \sum_j [ijP_d[i,j]] - \mu_i \mu_j}{\sigma_i \sigma_j}$$

Simple Visual Features

- Edge features:

```
gray_image = rgb2gray(my_image);
double_i = im2double(gray_image);
subplot(221),imshow(gray_image);
edge_image1 = edge(double_i,'sobel');
subplot(222),imshow(edge_image1);
edge_image2 = edge(double_i,'prewitt'
subplot(223),imshow(edge_image2);
edge_image3 = edge(double_i,'canny')
subplot(224),imshow(edge_image3);
```



Advanced Visual Features

- Shape features
  - A good shape representation feature for an object should be invariant to translation, rotation, and scaling
  - The use of shape features for image retrieval has been limited to special applications where objects or regions are readily available
  - Shape description can be categorized into either boundary-based or region-based method

Advanced Visual Features

- Middle-level features
  - Face detection:
    number of faces, location of face, etc.
  - Region detection:
    types of shapes: polygon, triangle, and circle, etc.
  - Categories:
    indoor and outdoor, play and non-play, etc.

Advanced Video Features

- High-level features:
- Objects, concepts, events, etc.
- Existence of an entity:

  e.g., trees
- Descriptive meaning

  e.g., sports

Traditional Approach (2012)

- Feature Selection
  - Choose a feature subset from the original feature set, which best represents the target semantic concepts
  - Having more features should surely result in more discriminating power, but adding irrelevant or distracting features often confuses system
- Algorithms
  - Filter model, Wrapper Model, Hybrid Model
  - Supervised feature selection, Unsupervised feature selection