Lecture Outline

- The content-based multimedia retrieval problem
- System components
- Describing multimedia content
  - Image features
    - Color, Texture, others
  - Audio and text features
  - Video representation and description → next week by Eric
- Distance metrics
- Evaluation of retrieval systems

- Guidelines for paper reading and presentation
- You questions: color quantization, perception vs. indexing
Information Retrieval Systems

- Conventional (library catalog).
  Search by keyword, title, author, etc.

- Text-based (Lexis-Nexis, Google, Yahoo!).
  Search by keywords. Limited search using queries in natural language.

- Multimedia (QBIC, WebSeek, SaFe)
  Search and matching by perceptual appearance

- Question answering systems (Ask, NSIR, Answerbus)
  Search in (restricted) natural language

- Clustering systems (Vivisimo, Clusty)

- Research systems (Lemur, Nutch)

Courtesy of Dragomir Radev, COMS 6998 Search Engine Technology
Why Visual Description and search?

Google/YouTube — “Basketball”
- Scope of results
  - Does not broadly search the Web
  - Includes “The Office” in Top 5
  - Cannot distinguish matches showing basketball

SearchVideo (Blinkx) — “Basketball”
- Scope of results
  - 214,000 matches related to “basketball”
  - No way to limit results to relevant scenes showing basketball games

- Manual annotation is tedious, insufficient, and inaccurate
- Computers cannot understand images
- Comparison of visual features enables comparison of visual scenes
- Visual search will create tools for organizing filtering and searching through large amounts of visual data
What functionalities should each component have? What are the bottlenecks of the system?

Dan Russell (Google), Searching for the mind of the searcher, Keynote speech, Joint Conference on Digital Libraries (JCDL), Vancouver, Canada, June 20, 2007

Tao Yang (Ask.com), Large-Scale Internet Search, Keynote speech, IEEE Intl. Conf. on Multimedia, Beijing, China, July 2007
Content-based Multimedia Information Retrieval System

- Description and Indexing: how do we describe and computationally represent text, images, audio, video etc.?
- Matching: how to match one description of multimedia with another?
- "Multimodal" fusion: how do we leverage different information sources, such as link, keyword, image descriptors, user tags ...
Describing an Image

What do we use?
- Pixels, features, metadata ...

Recognition rate > 99%

Desired Properties of Visual Features

- Invariance:
  - Rotation, scaling, cropping, shift, etc.

- Subjective relevance
  - illumination, object pose, background clutter, occlusion, intra-class appearance, viewpoint
Desired Properties of Visual Features

- **Invariance:**
  - Rotation, scaling, cropping, shift, etc.

- **Subjective relevance**
  - Illumination, object pose, background clutter, occlusion, intra-class appearance, viewpoint

- **Effective representation**
  - Low dimensionality; distance metric well-defined; efficient to compute.

\[
\begin{bmatrix}
    x_1, x_2, \ldots, x_{10}, \ldots, x_{100}, \ldots, x_{10^6}
\end{bmatrix}
\]

\[
\begin{bmatrix}
    x_1, x_2, \ldots, x_{10}
\end{bmatrix}
\]

\[d(\text{green}, \text{gray})=?]\]

\[\begin{bmatrix}
    0.001s, 1s, 100s?
\end{bmatrix}\]
Describing an Image

- What visual features?

- What is available in the data?
- What features does the human visual system (HVS) use?

- Color: beyond cats-and-dogs vision
- Texture: visual patterns, surface properties, cues for depth
- Shape: boundaries and measurement of real world objects and edges
- Motion: camera motion, object motion, depth from motion
General Approach for Visual Indexing

- Fundamental approach is from pattern recognition work
  - Group pixels, process the group and generate a feature vector
  - Summarize, if necessary, features over an entire image or unit of comparison (image patch, region, blob)
  - Discrimination via (transform and ) feature vector distance
  - Multidimensional indexing of the feature vectors

- This lecture: Do this for color and texture
  - Build a content-based image retrieval system
Lecture Outline

- The content-based multimedia retrieval problem
- System components
- Describing multimedia content
  - Image features
    - Color, Texture, others
  - Audio and text features
  - Video representation and description
- Distance metric
- Evaluation of retrieval systems

- Guidelines for paper reading and presentation
- You questions: color quantization, perception vs. indexing
Why Does a Visual System Need Color?

- An incomplete list:
  - To tell what is edible
  - To distinguish material changes from shading changes
  - To group parts of one object together in a scene
  - To find people’s skin
  - Check whether someone’s appearance looks normal/healthy
  - To compress images
  - ... ...
Colors in Cartoon Spectrum

<table>
<thead>
<tr>
<th></th>
<th>Violet</th>
<th>Blue</th>
<th>Green</th>
<th>Yellow</th>
<th>Orange</th>
<th>Red</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelength in nanometers</td>
<td>400</td>
<td>500</td>
<td>600</td>
<td>700</td>
<td>800</td>
<td>900</td>
</tr>
</tbody>
</table>

- **Red**: 400-700 nm
- **Green**: 400-600 nm
- **Blue**: 400-600 nm
- **Cyan**: 400-500 nm
- **Magenta**: 600-700 nm
- **Yellow**: 600-700 nm

Courtesy of W. Freeman http://www.ai.mit.edu/courses/6.801/Fall2002/
Color Representation

- A weighted combination of stimuli at three principal wavelengths in the visible spectrum (form blue=400nm to red=700nm).

Examples:

\[ \lambda = 500\text{nm} \rightarrow (\beta, \gamma, \rho) = (20, 40, 20) \]
\[ B = 100 \rightarrow (\beta, \gamma, \rho) = (100, 5, 4) \]
\[ G = 100 \rightarrow (\beta, \gamma, \rho) = (0, 100, 75) \]
\[ R = 100 \rightarrow (\beta, \gamma, \rho) = (0, 0, 100) \]
Tri-stimulus Representation

\[ P_1(\lambda) \]
\[ P_2(\lambda) \]
\[ P_3(\lambda) \]

\[ \alpha_1 \]
\[ \alpha_2 \]
\[ \alpha_3 \]

HVS

Same Response \((\beta, \gamma, \rho)\)

E.g., use are R, G, B as primary colors \(P_1, P_2, P_3\)

Compute correct \(\alpha_1, \alpha_2, \alpha_3\) s.t. the response \((\beta, \gamma, \rho)\) are the same as those of original color.
**Color Spaces and Color Order Systems**

- **Color Spaces**
  - RGB – cube in Euclidean space
    
    \[
    r = \frac{R}{R + G + B} \quad g = \frac{G}{R + G + B} \quad b = \frac{B}{R + G + B}
    \]

- **Standard representation used in color displays**

- **Drawbacks**
  - RGB basis not related to human color judgments
  - Intensity should be one of the dimensions of color
  - Important perceptual components of color are hue, brightness and saturation
Color Spaces and Color Order Systems

- **HSI-cone (cylindrical coordinates)**

\[
\begin{bmatrix}
I \\
V_1 \\
V_2
\end{bmatrix} = \begin{bmatrix}
1/3 & 1/3 & 1/3 \\
-1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} \\
1/\sqrt{6} & -1/\sqrt{6} & 0
\end{bmatrix} \begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
\]

\[H = \tan^{-1}\left(\frac{V_2}{V_1}\right)\]

\[S = (V_1^2 + V_2^2)^{1/2}\]

- **Opponent-Cartesian**

\[
\begin{bmatrix}
R - G \\
Bl - Y \\
W - Bk
\end{bmatrix} = \begin{bmatrix}
1 & -2 & 1 \\
-1 & -1 & 2 \\
1 & 1 & 1
\end{bmatrix} \begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
\]

- **YIQ-NTSC television standard**

\[
\begin{bmatrix}
I \\
Q \\
Y
\end{bmatrix} = \begin{bmatrix}
0.6 & -0.28 & -0.32 \\
0.21 & -0.52 & 0.31 \\
0.3 & 0.59 & 0.11
\end{bmatrix} \begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
\]
Perceptual Representation Of HSI Space

- Brightness varies along the vertical axis
- Hue varies along the circumference
- Saturation varies along the radius
<table>
<thead>
<tr>
<th>Color coordinate system</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. C.I.E. spectral primary system: R, G, B</strong></td>
<td>Monochromatic primary sources $P_1$, red = 700 nm, $P_2$, green = 546.1 nm, $P_3$, blue = 435.8 nm. Reference white has flat spectrum and $R = G = B = 1$. See Figs. 3.13 and 3.14 for spectral matching curves and chromaticity diagram.</td>
</tr>
<tr>
<td><strong>2. C.I.E. $X, Y, Z$ system</strong></td>
<td>$Y = $ luminance</td>
</tr>
<tr>
<td>$Y = $ luminance</td>
<td>$\begin{bmatrix} X \ Y \ Z \end{bmatrix} = \begin{bmatrix} 0.490 &amp; 0.310 &amp; 0.200 \ 0.177 &amp; 0.813 &amp; 0.011 \ 0.000 &amp; 0.010 &amp; 0.990 \end{bmatrix} \begin{bmatrix} R \ G \ B \end{bmatrix}$</td>
</tr>
<tr>
<td><strong>3. C.I.E. uniform chromaticity scale (UCS) system: $u, v, Y$</strong></td>
<td>$u = \frac{4X}{X + 15Y + 3Z} = \frac{4x}{x + 12y + 3}$</td>
</tr>
<tr>
<td>$u, v = $ chromaticities</td>
<td>$v = \frac{6Y}{X + 15Y + 3Z} = \frac{6y}{x + 12y + 3}$</td>
</tr>
<tr>
<td>$Y = $ luminance</td>
<td>$U = \frac{2X}{3}, V = Y, W = \frac{-X + 3Y + Z}{2}$</td>
</tr>
<tr>
<td>$U, V, W = $ tristimulus values corresponding to $u, v, w$</td>
<td>$U^* = 13W^*(u - u_0)$</td>
</tr>
<tr>
<td><strong>4. $U^<em>, V^</em>, W^*$ system (modified UCS system)</strong></td>
<td>$V^* = 13W^*(v - v_0)$</td>
</tr>
<tr>
<td>$Y = $ luminance $[0.01, 1]$</td>
<td>$W^* = 25(100Y)^{0.3} - 17, 1 \leq 100Y \leq 100$</td>
</tr>
<tr>
<td>$u_0, v_0 = $ chromaticities of reference white</td>
<td>$W^* = $ contrast or brightness</td>
</tr>
</tbody>
</table>
### Color Coordinate Systems (cont.)

5. \( S, \theta, W^* \) system:
   \[
   S = [(U^*)^2 + (V^*)^2]^{1/2} = 13W^*[((u-u_0)^2 + (v-v_0)^2)]^{1/2}
   \]
   \[
   \theta = \tan^{-1}\left(\frac{V^*}{U^*}\right) = \tan^{-1}\left(\frac{(v-v_0)(u-u_0)}{(u-u_0)^2 + (v-v_0)^2}\right), 0 \leq \theta \leq 2\pi
   \]
   \( W^* = \text{brightness} \)

6. **NTSC receiver primary** system \( R_N, G_N, B_N \)
   Linear transformation of \( X, Y, Z \). Is based on television phosphor primaries. Reference white is illuminant \( C \) for which \( R_N = G_N = B_N = 1 \).
   \[
   \begin{bmatrix}
   R_N \\
   G_N \\
   B_N
   \end{bmatrix} = 
   \begin{bmatrix}
   1.910 & -0.533 & -0.288 \\
   -0.985 & 2.000 & -0.028 \\
   0.058 & -0.118 & 0.896
   \end{bmatrix} 
   \begin{bmatrix}
   X \\
   Y \\
   Z
   \end{bmatrix}
   \]

7. **NTSC transmission system**:
   \[
   Y = 0.299R_N + 0.587G_N + 0.114B_N
   \]
   \( Y = \text{luminance} \)
   \[
   I = 0.596R_N - 0.274G_N - 0.322B_N
   \]
   \( I, Q = \text{chrominances} \)
   \[
   Q = 0.211R_N - 0.523G_N + 0.312B_N
   \]

8. \( L^*, a^*, b^* \) system:
   \[
   L^* = 25\left(\frac{100Y}{Y_0}\right)^{1/3} - 16, 1 \leq 100Y \leq 100
   \]
   \( L^* = \text{brightness} \)
   \[
   a^* = 500\left[(\frac{X}{X_0})^{1/3} - (\frac{Y}{Y_0})^{1/3}\right]
   \]
   \( a^* = \text{red-green content} \)
   \[
   b^* = 200\left[(\frac{Y}{Y_0})^{1/3} - (\frac{Z}{Z_0})^{1/3}\right]
   \]
   \( b^* = \text{yellow-blue content} \)

\( X_0, Y_0, Z_0 = \text{tristimulus values of the reference white} \)
Color Space Quantization

- Why quantization: fewer numbers, less noise (?)
- How many colors to keep
  - IBM QBIC 1995 → 16M(RGB) → 4096 (RGB) → 64 (Munsell) colors
  - Columbia U. VisualSEEK 1996 → 16M (RGB) → 166 (HSV) colors
    - 16 hues, 3 Val, 3 Sat, 4 Gray
  - Stricker and Orengo 1995 (Similarity of Color Images)
    - 16M (RGB) → 16 hues, 4 val, 4 sat = 128(HSV) colors
    - 16M (RGB) → 8 hues, 2 val, 2 sat = 32 (HSV) colors

- Independent quantization
  - each color dimension is quantized independently
- Joint quantization
  - color dimensions are quantized jointly
Color Histogram

- Feature extraction from color images
  - Choose GOOD color space
  - Quantize color space to reduce number of colors
  - Represent image color content using color histogram
  - Feature vector IS the color histogram

\[ h_{RGB}[r, g, b] = \sum_{m} \sum_{n} \begin{cases} 
1 & \text{if } I_R[m,n] = r, I_G[m,n] = g, I_B[m,n] = b \\
0 & \text{otherwise} 
\end{cases} \]

A color histogram represents the distribution of colors where each histogram bin corresponds to a color is the quantized color space

Question: what image transforms is color histogram invariant to?
Color Histogram (cont.)

- Advantages of color histograms
  - Compact representation of color information
  - Global color distribution
  - Histogram distance metrics

- Disadvantages
  - High dimensionality
  - No information about spatial positions of colors
Color Histogram

- Boticelli, da Vinci, Monet, or Miro?
**Color moments**

- Is there a more compact representation than color histogram?
- Compute moment statistics in each color channel.

HW data: color moments in “Lab” color Space
http://en.wikipedia.org/wiki/Lab_color_space
Histograms of Partitioned Image

Divide image up into rectangles.
Compute separate histogram for each partition.

Rectangles can overlap.

http://www.ai.mit.edu/courses/6.801/Fall2002/
Retrieval by “color layout” in IBM’s QBIC

http://www.ai.mit.edu/courses/6.801/Fall2002/
Indexing with Color Correlograms
[Zabih, et al.]

**Problem:** Pictures taken from slightly different view positions can look substantially different with a color histogram similarity measure.

**Proposed solution:** Compute color co-occurrence statistics [Haralick 1979].

http://www.ai.mit.edu/courses/6.801/Fall2002/
Color Correlogram

[Zabih, et al.]

For each image, estimate the probability that a pixel of some color lies within a particular distance of pixel of another color.

http://www.ai.mit.edu/courses/6.801/Fall2002/
Estimating Color Correlogram

Consider set of distances of interest \([d=\{1,2,...,d\}]\)
Measure pixel distance with \(L_\infty\) norm.
Consider \(m\) possible colors \(c_i \in \{c_1, c_2, ..., c_m\}\).

Offline, for each image:
Construct a correlogram that has \(m \times m \times d\) bins, initialize=0.
For each pixel \(p_i\) in the image, find it’s color \(c_i\)
for each distance \(k \in \{1,2,...,d\}\)
for each pixel at distance \(k\) from \(p_i\)
increment bin \((i,j,k)\)
end
end
end
Normalize correlogram by a scale factor (see Zabih, et al.)

http://www.ai.mit.edu/courses/6.801/Fall2002/
Texture

What is texture?
- "stylized subelements, repeated in meaningful ways"
- May have quasi-stochastic macro structure (e.g. bricks), each with stochastic micro structure

Why texture?
- Application to satellite images, medical images
- Useful for describing and reproducing contents of real world images, i.e., clouds, fabrics, surfaces, wood, stone

Challenging issues
- Rotation and scale invariance (3D)
- Segmentation/extraction of texture regions from images
- Texture in noise
Texture descriptors

- Co-occurrence Matrix - (image with m levels)

\[
Q_{R(\theta, d)}(i, j) = \begin{pmatrix}
Q_{R(\theta, d)}(0,0) & \ldots & Q_{R(\theta, d)}(0,m) \\
\vdots & \ddots & \vdots \\
Q_{R(\theta, d)}(m,0) & \ldots & Q_{R(\theta, d)}(m,m)
\end{pmatrix}
\]

where,

\(R(\theta, d) = \) relation between two pixels, e.g., 'north', 'NW'

\(I[x_0, y_0] = i \) and \(I[x_1, y_1] = j \) and

\(y_1 = y_0 + d \sin(\theta)\) and \(x_1 = x_0 + d \cos(\theta)\)

- Popular early texture approach
Gray-level Co-occurrence Matrix

From matlab documentation for graycomatrix.m:

The following figure shows how graycomatrix calculates several values in the GLCM of the 4-by-5 image I. Element (1,1) in the GLCM contains the value 1 because there is only one instance in the image where two, horizontally adjacent pixels have the values 1 and 1. Element (1,2) in the GLCM contains the value 2 because there are two instances in the image where two, horizontally adjacent pixels have the values 1 and 2. graycomatrix continues this processing to fill in all the values in the GLCM.

```
glcms = graycomatrix(I, param1, val1, param2, val2,...) returns one or more gray-level co-occurrence matrices, depending on the values of the optional parameter/value pairs. Parameter names can be abbreviated, and case does not matter.
```

What is the GLCM for these 2x2 images with offset (1,0) and (1,1):

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Co-occurrence Matrix
(also called Grey-Level Dependence, SGLD)

- Measures on $Q_{R(\theta,d)}(i,j)$
  - Energy
    $$E(d, \theta) = \sum_i \sum_j Q_{R(\theta,d)}(i, j)$$
  - Entropy
    $$H(d, \theta) = \sum_i \sum_j \frac{Q_{R(\theta,d)}(i, j)}{E} \log(E/Q_R(i, j))$$
  - Correlation
    $$C(d, \theta) = \sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \sigma_y} \cdot Q_R(i, j)$$
  - Inertia
    $$I(d, \theta) = \sum_i \sum_j (i-j)^2 Q_R(i, j)$$
  - Local Homogeneity
    $$L(d, \theta) = \sum_i \sum_j \frac{1}{1+(i-j)^2} Q_R(i, j)$$

- Statistical Measures
- None corresponds to a visual component.
Gray-scale Co-occurrence Stats of Paintings

Contrast

Correlation

Energy

Homogeniety
Filter approaches for texture description

- Fourier Domain Energy Distribution
  - Angular features (directionality)
    \[ V_{\theta_1,\theta_2} = \iint |F(u, v)|^2 \, du \, dv \]
    where
    \[ \theta_1 \leq \tan^{-1}\left(\frac{v}{u}\right) \leq \theta_2 \]
  - Radial features (coarseness)
    \[ V_{r_1,r_2} = \iint |F(u, v)|^2 \, du \, dv \]
    where
    \[ r_1 \leq u^2 + v^2 < r_2 \]
Gabor filters

- Gaussian windowed Fourier Transform
  - Make convolution kernels from product of Fourier basis images and Gaussians

\[
\text{2D DFT basis} \times \text{Gaussian window} = \text{2D Gabor filters}
\]

Odd (sin) | Even (cos)
--- | ---

Frequency
Example: Filter Responses

Filter bank

Input image

from Forsyth & Ponce
Laws Filters [1980]

- Non-Fourier type bass
- Matched better to intuitive texture features
- Examples of filters (out of total 12)

\[
\begin{bmatrix}
-1 & -4 & -6 & -4 & -1 \\
-2 & -8 & -12 & -8 & 2 \\
0 & 0 & 0 & 0 & 0 \\
2 & 8 & 12 & 8 & 2 \\
1 & 4 & 6 & 4 & 1
\end{bmatrix}
\begin{bmatrix}
-1 & 0 & 2 & 0 & -1 \\
-2 & 0 & 4 & 0 & -2 \\
0 & 0 & 0 & 0 & 0 \\
2 & 0 & -4 & 0 & 2 \\
1 & 0 & 2 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & -4 & 6 & -4 & 1 \\
-4 & 16 & -24 & 16 & -4 \\
6 & -24 & 36 & -24 & 6 \\
-4 & 16 & -24 & 16 & -4 \\
1 & -4 & 6 & -4 & 1
\end{bmatrix}
\]

- Measure energy of output from each filter
Tamura Texture

- Methods for approximating intuitive texture features
- Example: ‘Coarseness’, others: ‘contrast’, ‘directionality’
  - Step 1: Compute averages at different scales, 1x1, 3x3, 5x5 pixels
    \[ \forall (x, y), \ A_k(x, y) = \sum_{i=x-2^{k-1}}^{x+2^{k-1}} \sum_{j=y-2^{k-1}}^{y+2^{k-1}} \frac{f(i, j)}{(2^k + 1)^2} \]
  - Step 2: compute neighborhood difference at each scale
    \[ \forall (x, y), \ E_{k,h}(x, y) = |A_k(x + 2^{k-1}, y) - A_k(x - 2^{k-1}, y)| \]
  - Step 3: select the scale with the largest variation
    \[ \forall (x, y) \text{ determine } E_k = \max(E_1, E_2, \ldots, E_L), \quad S_{Best}(x, y) = 2^k \]
  - Step 4: compute the coarseness
    \[ F_{CRS} = \frac{1}{MN} \sum_{j=1}^{m} \sum_{i=1}^{n} S_{Best}(i, j) \]
A wide range of filters for textures


- Tamura Texture, Zernike moments, Steerable filters
- Ring/wedge filters, dyadic Gabor filter banks, wavelet transforms, wavelet packets and wavelet frames, quadrature mirror filters, discrete cosine transform, eigenfilters, optimized Gabor filters, linear predictors, optimized finite impulse response filters ...
Shape

- Needs known or pre-segmented shape
- Applicable to trademark, engineering drawings or 3-D models
- Descriptors
  - Area, perimeter, elongation, eccentricity, moments, Fourier descriptors, chain codes, reflective symmetry descriptors ...

http://amp.ece.cmu.edu/projects/TrademarkRetrieval/
http://www.cs.princeton.edu/gfx/proj/shape/
Audio features

- Speech, music, or environmental sound?
- Waveform: zero-crossing
- Spectrum: centroid, spread, flatness/entropy, cepstra, MFCC, ...
- Harmonicity: degree, stability
- Pitch, attack time, melody structure ...

Sound is essentially “visible” – use image analysis to index sound?

Female singing  Female and mixed speech
Distance Metrics between Vectors

- **L1 distance**
  \[ D_{L1}(x_1, x_2) = \sum_i |x_1(i) - x_2(i)| \]

- **L2 distance**
  \[ D_{L2}^2(x_1, x_2) = \sum_i |x_1(i) - x_2(i)|^2 \]

- **Histogram Intersection**
  \[ D_I = 1 - \frac{\sum_i \min\{x_1(i), x_2(i)\}}{\min\{\sum_i x_1(i), \sum_i x_2(i)\}} \]

- **Mohalanobis distance**
  \[ D_{mah}^2 = (x_1 - x_2)^T C_x^{-1} (x_1 - x_2) \]
  where \( C_x \) is the covariance matrix
Similarity between Sets

\[ [x_1, x_2 \ldots x_{10} \ldots x_{100} \ldots x_{106}] \]

vs

\[ [y_1, y_2 \ldots y_{10} \ldots y_{100} \ldots y_{106}] \]

- Video = sets of frames
- Image = sets of pixels
- Image patch = sets of descriptors
- Speech/text = sets of words
- ...
Earth Mover’s Distance (EMD)

- Rubner, Tomasi, Guibas ‘98
- Mallow’s distance in statistics in 1950’s
- Transportation Problem [Dantzig’51]

$I$: set of suppliers
$J$: set of consumers
$c_{ij}$: cost of shipping a unit of supply from $i$ to $j$

Problem: find the optimal set of flows $f_{ij}$ to

$$\begin{align*}
\text{minimize} & \quad \sum_{i \in I} \sum_{j \in J} c_{ij} f_{ij} \\
\text{subject to} & \quad f_{ij} \geq 0, i \in I, j \in J \quad (\text{No reverse shipping}) \\
& \quad \sum_{i \in I} f_{ij} = y_j, j \in J \quad (\text{satisfy each consumer need/capacity}) \\
& \quad \sum_{j \in J} f_{ij} \leq x_i, i \in I \quad (\text{bounded by each supplier’s limit}) \\
& \quad \sum_{j \in J} y_j \leq \sum_{i \in I} x_i \quad (\text{feasibility})
\end{align*}$$
Advantage of EMD

- Efficient implementations exist (Simplex Method)
- Also support partial matching ($||I|| < ||J||$, e.g., histogram defined in different color spaces, or scales)
- If the mass of two distributions equal, then EMD is a true metric
- Allow flexible representation, e.g., matching multiple regions in each image
  - Multiple region in one image, each region represented by individual feature vector

Region set: $\{R_1, R_2, R_3\}$
Region set: $\{R_1', R_2', R_3', R_4'\}$

$C_{ij} = \text{dist}(R_i, R_j')$, which can be based on EMD also
EMD of Color Histogram

\[ h = [h(1), h(2), \ldots, h(M)], \quad g = [g(1), g(2), \ldots, h(N)], \quad \text{assume } \sum_{j} g(j) \leq \sum_{i} h(i) \]

\[
EMD(h, g) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} C_{ij} f_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N} f_{ij}}
\]

\[
= \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} C_{ij} f_{ij}}{\sum_{j=1}^{N} g_j} \quad \text{Fill up each hole}
\]

\( C_{ij} \): distance between color \( i \) in color space \( h \) and color \( j \) in color space \( g \)

\( f_{ij} \): move \( f_{ij} \) units of mass from color \( i \) in \( h \) to color \( j \) in \( g \)

- Normalization by the denominator term
  - Avoid bias toward low mass distributions (i.e., small images)
  - what’s the difference if both \( h \) and \( g \) are normalized first?
    - exact matching of sub-parts is changed.
Question
- When to use which metric?
- Can various metrics be used together?
- How do we combine them?
Evaluation

N Images in DB  \[ V_n = 1 \] "Relevant"
K ranked returned Result  \[ 0 \] "Irrelevant"  \( n = 0 \cdots N - 1 \)

- Recall  \[ R = A/(A + C) \]
- Precision  \[ P = A/(A + B) \]
- Fallout  \[ F = B/(B + D) \]
- Combined  \[ F_1 = \frac{P \cdot R}{{(P + R)}/2} \]

- Detection  \[ A = \sum_{n=0}^{K-1} V_n \]
- False Alarms  \[ B = \sum_{n=0}^{K-1} (1 - V_n) \]
- Misses  \[ C = (\sum_{n=0}^{N-1} V_n) - A \]
- Correct Dismissals  \[ D = (\sum_{n=0}^{N-1} (1 - V_n)) - B \]
Evaluation Measures

Precision Recall Curve

2. Receiver Operating Characteristic (ROC Curve)

3. Relative Operating Characteristic

4. Precision at depth K

5. 3-point P value

\[ P_k \text{ at cut off } k = \text{int}(\sum_{n=0}^{N-1} V_n) \]

\[ \text{Avg } P \text{ at } R = 0.2 \quad 0.5 \quad 0.8 \]
Evaluation Metric: Average Precision

S → Ranked list of data in response to a query

\[
\begin{align*}
D_{15} & \quad D_8 & \quad D_{63} & \quad D_{21} & \quad \ldots & \quad \ldots & \quad D_s \\
\text{Ground truth} & \quad 1 & \quad 0 & \quad 1 & \quad 1 & \quad 0 & \quad 0 & \quad 0 \\
\text{Precision} & \quad 1/1 & \quad 1/2 & \quad 2/3 & \quad 3/4 & \quad 3/5 & \quad 3/6 & \quad 3/7
\end{align*}
\]

Average precision: \[ AP = \frac{1}{R} \sum_{j=1}^{s} \frac{R_j}{j}, \quad R: total \ number \ of \ relevant \ data \]

AP measures the average of precision values at R relevant data points
Evaluation Metric: Average Precision

- **Alternative Measure**
  - Ranked result are manually inspected to a depth of $N_1$
    - E.g., in TREC VIDEO 2003, $N_1 = 100$; in TREC VIDEO 2004, $N_1 = 1000$

- **Observations (AP)**
  - AP depends on the rankings of relevant data and the size of the relevant data set. E.g., $R=10$

  **Case I:**
  - Pre: $1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ AP=1$

  **Case II:**
  - Pre: $1/2 \ 1/2 \ 1/2 \ 1/2 \ 1/2 \ 1/2 \ 1/2 \ 1/2 \ 1/2 \ 1/2 \ 1/2 \ 1/2 \ 1/2 \ 1/2 \ AP=1/2$

  **Case III:**
  - Pre: $1/11 \ 2/12 \ ... \ ... \ 10/20 \ AP \sim \ 0.3$
Evaluation Metric: Average Precision

**Observations (AP)**

- E.g., R=2

  **Case I:**
  
  \[
  + + - - - - 
  \]

  AP=1

  **Case II:**
  
  \[
  - - - - + - - - - + 
  \]

  Precision 0.2 0.2

  AP=0.2

- AP is different from interpolated average of precision values
1 Background

As the number of images and videos continues to increase, we will see more intelligent forms of image search applications develop. In this homework assignment, you will create a basic system that performs content-based image retrieval (CBIR). You must rank a set of images given a single query image. This homework will expose you to three essential tasks for indexing and searching video: feature extraction, distance metric choice, and performance analysis. You will be provided with skeleton sample code developed in Matlab and there are two opportunities to obtain bonus points towards your final course grade.

1.1 Dataset & Feature Extraction

This assignment uses a subset of images derived from a work that analyzes the performance of different low-level features for automatically annotating consumer photos\(^1\). This image set is derived from downloads from Flickr\(^2\) and Yahoo\(^3\) so it should acclimate you to the challenges of an image search system. Your CBIR system will be searching the dataset for the best match to a small number of query images; dataset examples are shown in Figure 1.

![Dataset Examples](image)

Figure 1. Example consumer photos of famous locations: the White House, the Brooklyn Bridge, Mount Rushmore, and the Pyramids of Giza.

Feature extraction is the process of analyzing and computing numerical representations of an image. Common low-level features used in the image processing community are color moments, edge detection histograms, Gabor or wavelet texture, and shape information.

To expedite system development, we have pre-computed low-level color moments and texture features and provide these files to the CourseWorks system. Both feature sets are formatted in a simple space delimited format (shown below), so you can easily import these into any programming environment of your choice.

```plaintext
<file name> <feature1> <feature2> <feature3> ... <feature N>
...
<file name N> <feature1> <feature2> <feature3> ... <feature N>
```

Figure 2. Example feature format for pre-computed features.

1.1.1 Dataset description
3 Grading Checkpoints

As part of the research process, individuals are required to rigorously discuss their assertions and insights about their findings. In a write-up in Microsoft Word or PDF format, please address each of the checkpoints outlined below.

- Discussion (written with supporting graphs if necessary)
  
  1. Inspect the images that you downloaded. What common trends or patterns do you observe that might be exploited among image features? (1 point)
  
  2. Derive two new features to process the images and analyze the performance of this feature. Only one of your features can be a form of normalizing an existing pre-computed feature. How does your feature capitalize on data that others did not? How can you leverage your feature in a large dataset (i.e. build a codebook or optimize its extraction). (1 point)
  
  3. Visually inspect results of your system. Without looking at the performance, what are your impressions of the returned results? Which features did you expect to perform and did they do so? Can you learn anything by also looking at the worst matches? Why or why not? (1 point)
  
  4. Now, also using empirical results as your evidence, discuss the strengths and weaknesses of the distance metrics used; at least the L1 metric (described above) and one other chosen by you. Why did you choose this secondary metric? (2 points)
  
  5. Inspecting your performance graphs, what does this indicate about CDIR operations at different depths? What do you expect to happen if the size (number of images) was increased in terms of inter-class confusion and feature discriminability? (2 points)

- Data files
  
  1. Graphical plots (like figure 3) of different CRIR permutations: at least two distance metrics (only

  2. Your complete source code for this assignment with adequate documentation. (2 points)

- Bonus (optional)
  
  1. Use the features for the test set and run your algorithm on the 40 query images in this text file. The best performance from the entire class will win these points. (+2 points)

  2. Create a system that performs multi-modal fusion of scores that performs better or roughly equal to any single metric or modality alone. Clearly describe your reasoning for choosing this fusion system and explicitly document the formula or algorithm you used along with graphs of its performance like figure 3. (+1 point)
Paper Review and Demo

- **Review**
  - Background review and examples
  - Problem addressed and main ideas
  - Insights about why it works
  - Limitation, generality, and repeatability
  - Alternatives and comparisons

- **Demo**
  - Are the software and data available and repeatable?
  - Reconstruct the method and try on toy data set? (from some available generic toolkit)
  - Analysis of results (not just accuracy numbers, offer explanations and verifiable theories about observations)
  - Demo code archived on class site and shared with others
Paper review and demo

Sample work schedule

- Week 1: understand paper, review and research
- Week 2: simulate a toy problem using available data set and tools
- Week 3: prepare presentation

Each student discusses paper and demos with Eric, Prof. Chang or me two weeks before class.

- Upload the slide and codes to CourseWorks before class
- Presentation
  - 30 mins each paper (including demo)
  - We will provide additional materials about the subject.
Choose your paper/topic now!

- Four weeks of paper presentation
- 4 papers each week
  ... starting three weeks from now!

- Email us of your choices by Tuesday!

<table>
<thead>
<tr>
<th>10/8/07</th>
<th>Paper Presentation</th>
<th>Students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10/29/07</th>
<th>Paper Presentation</th>
<th>Students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Checker-shadow illusion:
The squares marked A and B are the same shade of gray.

Edward H. Adelson
There are object-specific response areas in our brain. The responses from movies are significantly different from those from still images.

[Hasson et al., Science’04]