CBIVR: Content-Based Image and Video Retrieval

Prepared by Stan Sclaroff
(with a few slides from Linda Shapiro)
for 6.801/6.866
December 3, 2002
QBIC: Query by Image Content

- First commercial system
- Search by:
  - color percentages
  - color layout
  - texture
  - shape/location
  - keywords

Try their demo: http://www.qbic.almaden.ibm.com
Image Features / Distance Measures

User

Image Database

Query Image

Retrieved Images

Images

Image Feature Extraction

Feature (vector) Space

Distance Measure
Query Formulation Methods

• QBE: Query by Example
  – Positive and negative examples
• Text description
• Query by sketch
• Cluster-based retrieval
• Relevance feedback
Query by Sketch

Example taken from Jacobs, Finkelstein, & Salesin
Fast Multi-Resolution Image Querying, SIGGRAPH 1995
Relevance Feedback

Application Areas

- Images and video on the web
- *Igrep*: Images and video in email and local files
- Individual collections of video or family photos
- Military intelligence, homeland security
- Archives: stock photos, stock film/video footage
- Access to museum collections
- Trademark and copyright infringement
- Medical information systems
CBIVR: Some Key Issues

Searching a large database for images or video clips that match a query:

– What kinds of databases?
– What kinds of queries?
– What constitutes a match?
– How do we make such searches efficient?
– How to quantitatively evaluate performance?
CBIVR: Some Key Issues

Searching a large database for images or video clips that match a query:

– What kinds of databases?
– What kinds of queries?
– What constitutes a match?
– How to make such searches efficient?
– How to quantitatively evaluate performance?
Quantitative Evaluation of CBIVR Performance
A Standard Information Retrieval Evaluation Measure

For a given query \( q \):

- \( R_a \) = set of relevant documents in answer set \( A \)
- \( R \) = set of relevant documents for \( q \)

\[
\text{Recall} = \frac{|R_a|}{|R|} \quad \text{Precision} = \frac{|R_a|}{|A|}
\]

Ideally these values are close to one.
A Standard Information Retrieval Evaluation Measure

For a given query $q$:
- $R_a = \text{set of relevant documents in answer set } A$
- $R = \text{set of relevant documents for } q$

**Recall** = \[
\frac{|R_a|}{|R|}
\]

**Precision** = \[
\frac{|R_a|}{|A|}
\]

**Problem**: What is relevant?
The relevance judgments of competent human informants can differ.
Generally, as the recall level rises, the level of precision falls.
Which system is better?
Desirable Precision vs. Recall is Application-Dependent!

- Images and video on the web
- *Igrep*: Images and video in email and local files
- Individual collections of video or family photos
- Military intelligence, homeland security
- Archives: stock photos, stock film/video footage
- Access to museum collections
- Trademark and copyright infringement
- Medical information systems
Example Application:

CBIVR for the WWW
CBIVR for the WWW?

- Very large, unstructured database
- Diverse content
- No single standard image nor video format
- No standard illumination
- Images/video can be altered in Photoshop, etc.
- Lossy compression, color quantization, scanned

What are the precision and recall requirements?
Enter relevant keywords for images to search for:

mountain bike race

Number of returned images: 50

Search  Reset form
Select relevant images to guide search.
Select relevant images to guide search.
Google SafeSearch Filtering

Global Preferences (changes apply to all Google services)

Interface Language
Display Google tips and messages in: English
If you do not find your native language in the pulldown above, you can help Google create it through our Google in Your Language program.

Search Language
- Search for pages written in any language (Recommended).
- Search only for pages written in these language(s):
  - Arabic
  - Bulgarian
  - Catalan
  - Chinese (Simplified)
  - Chinese (Traditional)
  - Croatian
  - Czech
  - Danish
  - Dutch
  - English
  - Estonian
  - Finnish
  - French
  - German
  - Greek
  - Hebrew
  - Hungarian
  - Icelandic
  - Indonesian
  - Italian
  - Japanese
  - Lithuanian
  - Latvian
  - Luxembourgish
  - Macedonian
  - Malay
  - Maltese
  - Norwegian
  - Polish
  - Portuguese
  - Russian
  - Romanian
  - Russian
  - Serbian
  - Slovak
  - Slovenian
  - Spanish
  - Swedish
  - Turkish

SafeSearch Filtering
Google’s SafeSearch blocks web pages containing explicit sexual content from appearing in search results.
- Use strict filtering (Filter both explicit text and explicit images)
- Use moderate filtering (Filter explicit images only - default behavior)
- Do not filter my search results.
Finding Naked People
[Fleck, Forsyth, and Bregler 1996]

- Convert RGB color to HIS color
- Use the intensity component to compute a texture map
  \[ \text{texture} = \text{med2}\left( | I - \text{med1}(I) | \right) \]
- If a pixel falls into either of the following ranges, it’s a potential skin pixel
  \[
  \begin{align*}
  \text{texture} &< 5, \ 110 < \text{hue} < 150, \ 20 < \text{saturation} < 60 \\
  \text{texture} &< 5, \ 130 < \text{hue} < 170, \ 30 < \text{saturation} < 130
  \end{align*}
  \]

Look for LARGE areas that satisfy this to identify pornography.
Use simple grouping rules for limbs/trunk/legs (see their paper).
Image Representations for CBIR
Image Representations

There are roughly three levels of image representation used for CBIR:

1. Iconic – exact pixel values
2. Compositional – overall image appearance
3. Objects – things depicted in the image, their properties, and their relationships
Iconic Matching

Example applications:
- Copyright and trademark protection
- Duplicate removal
- Linking images used in evidence, for example child pornography

Problems in finding “exact” matches:
- Lossy compression, image scanning
- Color space conversion
- Photoshop-style transforms: blur, scale, rotate, warp, crop, cut, etc.
Iconic Matching

From Jacobs, Finkelstein, & Salesin
Fast Multi-Resolution Image Querying, SIGGRAPH 1995
Summary Representations of a Whole Picture: Color Histograms

[Swain and Ballard, IJCV 1991]
Color Histograms

Off-line, for each image
   create histogram with a bin for each color
   initialize each bin counter = 0
   for each pixel in image:
      increment bin counter corresponding to pixel color
   end

On-line, use histograms in image similarity measure:
   Euclidean, dot product, histogram intersection, etc.
The QBIC color histogram distance is:

\[ d_{\text{hist}}(I, Q) = (h(I) - h(Q))^T A (h(I) - h(Q)) \]

- \( h(I) \) is a K-bin histogram of a database image
- \( h(Q) \) is a K-bin histogram of the query image
- \( A \) is a K x K similarity matrix
### Similarity Matrix: $A$

$$
\begin{array}{cccccc}
R & G & B & Y & C & V \\
\hline
R & 1 & 0 & 0 & .5 & 0 & .5 \\
G & 0 & 1 & 0 & .5 & .5 & 0 \\
B & 0 & 0 & 1 & \text{?} & & \\
Y & \text{?} & & & 1 & & \\
C & \text{?} & & & & 1 & \\
V & & & & & & 1 \\
\end{array}
$$

How similar is blue to cyan?
Images Classified as Sunsets using Overall Color Histograms
Histograms of Partitioned Image

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

Rectangles can overlap.
Retrieval by “color layout” in IBM’s QBIC

<table>
<thead>
<tr>
<th>Images 1-8 out of 41</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image 1" /></td>
<td><img src="image2" alt="Image 2" /></td>
</tr>
<tr>
<td><img src="image3" alt="Image 3" /></td>
<td><img src="image4" alt="Image 4" /></td>
</tr>
<tr>
<td><img src="image5" alt="Image 5" /></td>
<td><img src="image6" alt="Image 6" /></td>
</tr>
<tr>
<td><img src="image7" alt="Image 7" /></td>
<td><img src="image8" alt="Image 8" /></td>
</tr>
</tbody>
</table>

Columns:  Rows:
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].
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.
Estimating Color Correlogram

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

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 $k \in \{1,2,\ldots,d\}$
    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.)
Similarity Measure

Given correlograms of query and target images, $\gamma(Q)$ and $\gamma(T)$, define similarity:

$$| Q - T |_{\gamma, L_1} \overset{\Delta}{=} \sum_{i,j \in [m], k \in [d]} | \gamma_{c_i c_j}^{(k)}(Q) - \gamma_{c_i c_j}^{(k)}(T) |$$

Improved measure includes normalization:

$$| Q - T |_{\gamma, d_1} \overset{\Delta}{=} \sum_{i,j \in [m], k \in [d]} \left( \frac{| \gamma_{c_i c_j}^{(k)}(Q) - \gamma_{c_i c_j}^{(k)}(T) |}{1 + \gamma_{c_i c_j}^{(k)}(Q) + \gamma_{c_i c_j}^{(k)}(T)} \right)$$
Earth Mover’s Distance (EMD)  
[Rubner, Guibas, & Tomasi 1998]

For each image, compute color signature:

Define distance between two color signatures to be the minimum amount of “work” needed to transform one signature into another.
Computing the Image Color Signature for EMD

- Transform pixel colors into CIE-LAB color space.
- Each pixel of the image constitutes a point in this color space.
- Cluster the pixels in color space, \( k-d \) tree based algorithm. Clusters constrained to not exceed 30 units in L,a,b axes.
- Find centroids of each cluster.
- Each cluster contributes a pair \( (p, w_p) \) to the signature
  - \( p \) is the average color.
  - \( w_p \) is the fraction of pixels in that cluster.

Typically there are 8 to 12 clusters.
EMD of Color Signatures

The \textit{work} needed to move a point, or a fraction of a point, to a new location is the portion of weight being moved, multiplied by the Euclidean distance between the old and new location.

Allow the weight of a single source to be partitioned among several destination points, and vice versa.

Can be solved with linear programming (see Rubner, et al.).
Multiscale EMD Formulation

Figure 1. Scale Estimation. (a) pattern, image, and pattern scaled by the scale estimate $c^a$. (b),(d) pattern, image signatures. (c) pattern signature with weights scaled by $c^0$. (e) $\text{EMD}(x_0, (Y, cu))$ v. c.
Retrieval Example

1) 22.98
   13037.jpg
2) 24.03
   178093.jpg
3) 24.8
   84068.jpg
4) 24.92
   84075.jpg
5) 25.8
   132098.jpg

6) 26.5
   84079.jpg
7) 26.66
   13033.jpg
8) 26.70
   51017.jpg
9) 26.81
   13151.jpg
10) 27.3
    13035.jpg
Example EMD Retrieval Results

Figure 3: The top ten images for a query that asked for 20% blue and 80% don’t care. (a) Traditional display. (b) MDS map.
Visualizing Dataset with EMD and Multidimensional Scaling
Orientation Histograms

Determine local orientation and magnitude at each pixel.

Multiscale Orientation Histograms

Off-line, for each image:

1. Compute steerable pyramid
2. For each level in the pyramid:
   a. For each pixel:
      i. Estimate local orientation and magnitude
      ii. If magnitude > threshold:
         1. Increment appropriate histogram bin
      End
   End
End
3. Circular blur histogram
End
Fast Multi-Resolution Image Querying
[Jacobs, Finkelstein, & Salesin SIGGRAPH 1995]

Off-line:

For each image
  – Compute Haar wavelet decomposition
  – Store truncated coefficients (top 60 of largest magnitude in each color channel)
  – Quantize remaining coefficients to –1 or 1 (for negative or positive values)
  – Store the coefficients as the image signature

end
Similarity Measure

Let $Q$ and $T$ represent just a single channel of the wavelet decomposition of the query and target images. Define the similarity measure:

$$
\| Q, T \| = w_{0,0} |Q_{0,0} - T_{0,0}| + \sum_{i,j} w_{i,j} |\widetilde{Q}[i, j] - \widetilde{T}[i, j]|$

Only consider terms where query has non-zero wavelet coefficients.
Minka and Picard here?
Image Representations

There are roughly three levels of image representation used for CBIR:

1. Iconic – exact pixel values
2. Compositional – overall image appearance
3. Objects – things depicted in the image, their properties, and their relationships
Shape Properties: Projection Matching
[VisualSeek, Smith&Chang 1996]

In projection matching, the horizontal and vertical projections form a histogram.

Feature Vector
(0,4,1,3,2,0,0,4,3,2,1,0)

What are the weaknesses of this method? strengths?
Global Shape Properties: Tangent-Angle Histograms

Is this feature invariant to starting point?
Del Bimbo Elastic Shape Matching

Sketch-based query

retrieved images
Shape-based Search in Photobook

[Pentland, Picard, & Sclaroff 1994]
Coordinate in Deformable Prototype Space

[Sclaroff 1995]

candidate

10.1

protoype 1

5.7

protoype 2
sqrt(modal strain) from prototype rabbits
Problems in shape-based indexing

Many existing approaches assume

• segmentation is given, or…
• human operator circles object of interest, or…
• lack of clutter and shadows, or…
• objects are rigid, or…
• planar (2-D) shape models, or…
• models are known in advance
Deformable template-based region grouping
[Sclaroff&Liu, 2001]

2D deformable template is trained on training data for object class
Image Partitioning via Optimization
Model-based Region Splitting

• Detect candidates for splitting based on model fitting cost value and a specified threshold.
• Determine candidate cuts based on model and curvature extrema of the region group boundary.
Index trees: basic idea

- Off-line:
  - Generate deformed instances of the object class
  - Compute their shape feature vectors
  - Create hierarchical indexing structure
- On-line:
  - Compute the shape feature vector for a potential region group
  - Fetch the most similar model instance via comparing the shape feature vectors
Shape Index Trees

Candidate region group
Shape-Population Retrieval
[Liu and Sclaroff 2000]
Blobworld
[Belongie, et al. 1998]

- Images are segmented on color plus texture
- User selects a region of the query image
- System returns images with similar regions
- Works really well for tigers and zebras

Demo: http://elib.cs.berkeley.edu/photos/blobworld
Blobworld Region Segmentation

- 8D descriptor computed for each pixel:
  - color in L*a*b* space
  - 3 texture features at selected scale: anisotropy, polarity, and contrast
  - pixel position (x,y)

- Represent each image as mixture of Gaussians, estimated via EM algorithm.

- Resulting pixel memberships form a segmentation of the image (after connected components analysis, etc.)
Example Blobworld Segmentation
Example Blobworld Queries
The NeTra System

[Ma and Manjunath, 1999]

Retrieve by region color, texture, shape and position.

Demo: http://vision.ece.ucsb.edu/netra/Netra.html
The NeTra System Overview
Image Features

- Use vector quantization (VQ) to build code book for RGB color, given training set of images chosen from database.
- Fourier descriptor of region contour is used to represent shape. Similarity measure: Euclidean.
Example Segmentation Results

NeTra Indexing/Retrieval

- Color, texture, and shape are stored in separate index structures.
- Spatial location/size of regions represented by centroid and minimum bounding rectangle.
- Use quad-tree and/or R-trees to organize index for efficient queries.
Spatial Relationships

Example queries:

- Find all images where A is within 50 pixels of B.
- Find all images where A appears to the right of B.
- Find all images in where apples are on tables.

Appropriate indexing structures:

- Quad-trees
- R-trees, R*-trees
- K-d trees
- etc.
Quad-tree

Find all images where the centroid of A is within 50 pixels of B.
Hierarchical Minimum Bounding Rectangles

Find all images where A is within 50 pixels of B.
Region Relation Graph

- sky
- sand
- tiger grass
- above
- adjacent
- inside
- above above
- image
- abstract regions
Eigenfaces in Photobook
[Pentland, Picard, Sclaroff 1995]
Object Detection: Rowley’s Face Finder

1. convert to gray scale
2. normalize for lighting*
3. histogram equalization
4. apply neural net(s) trained on 16K images
Name-It [Satoh, Nakamura, Kanade 1999]

Video

Face Sequence
Extraction

Video Caption
Recognition

Name Extraction

Transcript

>>> TAKING RISKS FOR PEACE IS A
THREAT PRESIDENT CLINTON SAID
SHOULD APPLY FROM EAST TO
WEST...
THOSE SENTIMENTS FOUND...

...

>>> TO ALL OF YOU WHO ASKED ME
TO DO WHAT I COULD TO HELP PEACE
TAKE ROOT, I PLEDGE...

ONE WAY OR THE OTHER
NEXT PRESIDENT IS IN
THE PRESIDENTIAL RACE.

>> THE SPEAKER, EVEN THOUGH,
HIS NAME WASN'T ON THE BALLOT.

Dictionary
Thesaurus
Parser

Names

Face—Name Association
(Co-Occurrence Evaluation)

Face Detection
Face Tracking

Face Sequences

Text Detection
Character Recognition

Queries

Who is ?

Name-to-Face Retrieval

Results

Clinton

Who is Newt Gingrich ?

Face-to-Name Retrieval

Face Similarity
Inexact String Matching
Analysis Result Integration
Face Tracking and Frontal View Extraction

<table>
<thead>
<tr>
<th></th>
<th>start</th>
<th>end</th>
<th>frontal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td><img src="image1" alt="Start Image" /></td>
<td><img src="image2" alt="End Image" /></td>
<td><img src="image3" alt="Frontal Image" /></td>
</tr>
<tr>
<td>(b)</td>
<td><img src="image4" alt="Start Image" /></td>
<td><img src="image5" alt="End Image" /></td>
<td><img src="image6" alt="Frontal Image" /></td>
</tr>
<tr>
<td>(c)</td>
<td><img src="image7" alt="Start Image" /></td>
<td><img src="image8" alt="End Image" /></td>
<td><img src="image9" alt="Frontal Image" /></td>
</tr>
<tr>
<td>(d)</td>
<td><img src="image10" alt="Start Image" /></td>
<td><img src="image11" alt="End Image" /></td>
<td><img src="image12" alt="Frontal Image" /></td>
</tr>
</tbody>
</table>
OCR for Video Captions

Figure 9: Typical Video Caption
Content-based Video Retrieval
VideoQ: Query by Sketch
[S.F. Chang, et al.]

Demo: http://www.ctr.columbia.edu/videoq/
Shot Boundary Detection Methods

Assumption: shot boundaries are discontinuities in space-time.
  - Compare color and orientation histograms in adjacent frames
  - Motion (flow) analysis
  - Multimodal approach: video + audio track

Other applications:
  Keyframe extraction. Intelligent fast forward.

Problems:
  - Transitions like wipe, fade, cross-dissolve
  - Camera motions: pan, zoom, etc.
  - Moving objects occupy large percentage of image
Video Skimming (CMU Informedia)
Event Detection, Indexing and Retrieval

Assign semantic labels to significant events in video:
- Explosion, car crash, door slam (audio/video track)
- Marilyn Monroe enters scene
- Pele scores goal
- Jay Leno tells joke and then delivers punch-line
- Two people exchange a briefcase in park
- etc.

There are events that are “latent” in the video database that are not of interest now, but may become interesting later.
Relevance Feedback
Relevance Feedback

In real interactive CBIR systems, the user should be allowed to interact with the system to “refine” the results of a query until he/she is satisfied.

Relevance feedback work has been done by a number of research groups, e.g.:

• The Photobook Project (Media Lab, MIT)
• The Leiden Portrait Retrieval Project
• The ImageRover Project at Boston U.
• The MARS Project (Tom Huang’s group at Illinois)
• PicHunter (Cox, et al. at NEC)
Information Retrieval Model

• An IR model consists of:
  – a document model
  – a query model
  – a model for computing similarity between documents and the queries
• Term (keyword) weighting
• Relevance Feedback
Term weighting in Info Retrieval

• Term weight
  – assigning different weights for different keyword (terms) according their relative importance to the document

• define $w_{ik}$ to be the weight for term $t_k$, $k=1,2,...,N$, in the document $i$

• Target document $i$ can be represented as a weight vector in the term space

$$T_i = [w_{i1}, w_{i2}, ..., w_{iN}]$$
Term weighting

• The query \( Q \) also is a weight vector in the term space

\[
Q = [w_{q_1}, w_{q_2}, \ldots, w_{q_N}]
\]

• The similarity between \( T \) and \( Q \)

\[
Sim(T, Q) = \frac{T \cdot Q}{\|T\| \|Q\|}
\]
Using Relevance Feedback

- The CBIR system should automatically adjust the weights that were given by the user for the relevance of previously retrieved documents.

- Most systems use a statistical method for adjusting the weights.

What are the problems in applying the IR relevance feedback paradigm in image and video retrieval?
Clustering Images [Barnard & Forsyth 2001]

Cluster on text only.

Cluster on image features only.
Clustering Images

Two clusters obtained using both text and image segment features.
Image/Video Databases

• Since databases can be large, computational complexity is very important
• Spatial data structures can help
• Hierarchical data structures, clustering
• Multiple metric strategies
• Embeddings