multimedia retrieval

some slides courtesy

Jimmy Lin, University of Maryland
Dr. Ramesh R. Sarukkai, Yahoo! Search
outline

- images
- video
- speech
A Picture...
... is comprised of pixels
This is nothing new!

Seurat, Georges, A Sunday Afternoon on the Island of La Grande Jatte
The Semantic Gap

<table>
<thead>
<tr>
<th>Semantics</th>
<th>Wolf on Road with Snow on Roadside in Yosemite National Park, California on 24/1/2004 at 23:19:11 GMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object Labels</td>
<td>Wolf, Snow, road</td>
</tr>
<tr>
<td>Objects</td>
<td>Prototypical combinations of descriptors</td>
</tr>
<tr>
<td>Descriptors</td>
<td>Segmented blobs, Salient regions, Pixel-level histograms, Fourier descriptors, etc...</td>
</tr>
<tr>
<td>Raw Media</td>
<td>Images</td>
</tr>
</tbody>
</table>
The Semantic Gap

• Content-based retrieval often fails due to the gap between information extractable automatically from the visual data (feature-vectors) and the interpretation a user may have for the same data.
  ...typically between low level features and the image semantics

• The current hot topic in multimedia IR research
The Semantic Gap

Raw Media

This is what we have to work with

Image-level descriptors

Content descriptors

This is what we want

Semantic content

Photo of Yosemite valley showing El Capitan and Glacier Point with the Half Dome in the distance.

This is what we have to work with

Semantic content
Sub-image matching

- Given a query image, find the parent image in the database with which it matches, either as a whole or as a part.
- Give location information showing where in the parent the query is positioned.
- The images may be very high resolution.
- The query and target may be at different resolutions.
Example 1: Query
Example 1: Result

- Best matching image with sub-image identified

NB. Query is before restoration work, target is a restored image. Query and target image also differ in resolution.
Example 2: Query
Example 2: Result

Best match found, with sub-image identified
...Subsequent Best Matches

Retrieved results start from top-left to bottom right.
The IR Black Box

Query -> Representation Function -> Query Representation

Representation Function -> Comparison Function

Multimedia Objects

Representation Function -> Document Representation

Index

Hits
Recipe for Multimedia Retrieval

• Extract features
  – Low-level features: blobs, textures, color histograms
  – Textual annotations: captions, ASR, video OCR, human labels

• Match features
  – From “bag of words” to “bag of features”
Visual Features ...
Combination of Evidence

Multi-modal Query
- Text Aspect
  - Osama Bin Laden
- Audio Aspect
- Motion Aspect
- Image Aspect

Multiple Modality Video Collection Analysis
- Script Index
- Video OCR Index
- Production Metadata Index
- Audio Index
- Motion Object Index
- Image Index

Weighted Fusion of Search Results

Final Ranked List
New Algorithm for Similarity-Based Retrieval of Images

- Images in the database are stored as JPEG-compressed images
- The user submits a request for search-by-similarity by presenting the desired image.
- The algorithm calculates the DC coefficients of this image and creates the histogram of DC coefficients.
- The algorithm compares the DC histogram of the submitted image with the DC histograms of the stored images.
Histogram of DC Coefficients for the Image “Elephant”
Comparison of Histograms of DC Coefficients
Example of Similarity-Based Retrieval Using the DC Histograms
Similarity-Based Retrieval of Compressed Video

- Partitioning video into clips - video segmentation
- Key frame extraction
- Indexing and retrieval of key frames
DC Histogram Technique Applied for Video Partitioning
Example of Similarity-Based Retrieval of Key Frames Using DC Histograms
outline

- images
- video
- speech
Images and Video

• A digital image := a collection of pixels
  – Each pixel has a “color”

• Different types of pixels
  – Binary (1 bit): black/white
  – Grayscale (8 bits)
  – Color (3 colors, 8 bits each): red, green, blue

• A video is simply lots of images in rapid sequence
  – Each image is called a frame
  – Smooth motion requires about 24 frames/sec

• Compression is the key!
The Structure of Video

- Video
- Scenes
- Shots
- Frames
TREC For Video Retrieval?

- TREC Video Track (TRECVID)
  - Started in 2001
  - Goal is to investigate content-based retrieval from digital video
  - Focus on the shot as the unit of information retrieval (why?)

http://www-nlpir.nist.gov/projects/trecvid/

- Test Data Collection in 2004:
  - 74 hours of CNN Headline News, ABC World News Tonight, C-SPAN
## Searching Performance

### Table 1: Modality and MAP Scores

<table>
<thead>
<tr>
<th>Modality</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: ASR + Closed Captions (CC)</td>
<td>0.155</td>
</tr>
<tr>
<td>ASR + CC + Video OCR</td>
<td>0.177</td>
</tr>
<tr>
<td>ASR + CC + VOOCR + Image Similarity weighted by query type</td>
<td>0.198</td>
</tr>
<tr>
<td>ASR + CC + VOOCR + Image Similarity weighted by development set query results</td>
<td>0.207</td>
</tr>
<tr>
<td>ASR + CC + VOOCR + Image Similarity weighted by development set query results + Person X retrieval</td>
<td>0.218</td>
</tr>
</tbody>
</table>

ASR = automatic speech recognition
CC  = closed captions
VOOCR = video optical character recognition

Market Trends

• Broadband doubling over next 3-5 years
• Video enabled devices are emerging rapidly
• Emergence of mass internet audience
• Mainstream media moving to the Web
• International trends are similar
• Money Follows...
Market Trends (2005)

- How many of you are aware of video on the Web?
  - Large portion of online users
- How many have viewed a video on the Web in
  - The last 3 months?
    - 50%
  - The last 6 months?
  - Ever?
- Would you watch video on your devices (ipod/wireless)?
  - 1M downloads in 20 days (iPod)
- How many of you have produced video (personal or otherwise) recently?
  - Continuing to skyrocket with digital camera phones/devices
- How many of you have shared that with your friends/community? Would you have liked to?
  - Huge interest & adoption in viral communities
Market Trends

- Technology more media friendly
  - Storage costs plummeting (GB → TB)
  - CPU speed continuing to double (Moore’s law)
  - Increased bandwidth
  - Device support for media
  - Adding media to sites drives traffic
  - Web continues to propel scalable infrastructure for media products/communities
Video Search

Structured + Unstructured Data

Video Production

Content Features + Meta-Content

Meta-Content

Video Consumption /Community

Unstructured Data
Video Search

- Media Information Retrieval
  - Been around since the late 1970s
    - Text Based/DB
      - Issues: Manual Annotation, Subjectivity of Human Perception
    - Content Based
      - Color, texture, shape, face detection/recognition, speech transcriptions, motion, segmentation boundaries/shots
      - High Dimensionality
      - Limited success to date.

Citations:
"Image Retrieval: Current Techniques, Promising Directions, and Open Issues" [Rui et al 99]
Video Search

Active Research Area

Graph from "A new perspective on Visual Information Retrieval", Horst Eidenberger, 2004
Black: "Image Retrieval"; Grey:"Video Retrieval"; IEEE Digital Library
Video Search

Popular features/techniques:
- Color, Shape, Texture, Shape descriptors
- OCR, ASR
- A number of prototype or research products with small data sets
- More researched for visual queries
Video Search: Features

Texture: Autocorrelation; Wavelet transforms; Gabor Filters
Color: Color Moments; Color Histograms; Color Autocorrelograms
Shape: Edge Detectors; Moment invariants; Animate Vision; Marr; Finite Element Methods; Shape from Motion;
Segmentation: Scene segmentation; Scene Segmentation; Shot detection;
OCR: Modeling; Successful OCR deployments;
Face: Face Detection algorithms; Neural Networks; EigenFaces
ASR: Acoustic analysis; HMMS; N-grams; CSR; LVCSR; Domain Specific;

media IR systems
NIST Video TREC Starts
Web Media Search
# Video Search: Features

## Color
- Robust to background
- Independent of size, orientation
- Color Histogram [Swain & Ballard]
- “Sensitive to noise and sparse”- Cumulative Histograms [Stricker & Orgengo]
- Color Moments
- Color Sets: Map RGB Color space to Hue Saturation Value, & quantize [Smith, Chang]
- Color layout- local color features by dividing image into regions
- Color Autocorrelograms

## Texture
- One of the earliest Image features [Harlick et al 70s]
- Co-occurrence matrix
- Orientation and distance on gray-scale pixels
- Contrast, inverse deference moment, and entropy [Gotlieb & Kreyszig]
- Human visual texture properties: coarseness, contrast, directionality, likeliness, regularity and roughness [Tamura et al]
- Wavelet Transforms [90s]
- [Smith & Chang] extracted mean and variance from wavelet subbands
- Gabor Filters
- And so on

## Region Segmentation
- Partition image into regions
- Strong Segmentation: Object segmentation is difficult.
- Weak segmentation: Region segmentation based on some homogeneity criteria

## Scene Segmentation
- Shot detection, scene detection
- Look for changes in color, texture, brightness
- Context based scene segmentation applied to certain categories such as broadcast news
Video Search: Features

**Shape**
- Outer Boundary based vs. region based
- Fourier descriptors
- Moment invariants
- Finite Element Method (Stiffness matrix- how each point is connected to others; Eigen vectors of matrix)
- Turing function based (similar to Fourier descriptor) convex/concave polygons[Arkin et al]
- Wavelet transforms leverages multiresolution [Chuang & Kao]
- Chamfer matching for comparing 2 shapes (linear dimension rather than area)
- 3-D object representations using similar invariant features
- Well-known edge detection algorithms.

**Face**
- Face detection is highly reliable
  - Neural Networks [Rwoley]
  - Wavelet based histograms of facial features [Schneiderman]
- Face recognition for video is still a challenging problem.
  - EigenFaces: Extract eigenvectors and use as feature space

**OCR**
- OCR is fairly successful technology.
- Accurate, especially with good matching vocabularies.
- Script recognition still an open problem.

**ASR**
- Automatic speech recognition fairly accurate for medium to large vocabulary broadcast type data
- Large number of available speech vendors.
- Still open for free conversational speech in noisy conditions.
Video Search: Video TREC

• Overview:
  – Shot detection, story segmentation, semantic feature extraction, information retrieval
  – Corpora of documentaries, advertising films
  – Broadcast news added in 2003
  – Interactive and non-interactive tests
  – CBIR features
  – Speech transcribed (LIMSI)
  – OCR
Video Search

Structured + Unstructured Data

Traditional Media Search

Content Features + Meta-Content

Video Production

Unstructured Data

Web Video Search

Meta-Content

Video Consumption /Community
Opportunities

Example 1:

- User query “zorro”

- User 1 wants to see Zorro videos
- User 2 wants to see Legend of Zorro movie clips
- User 3 just wants to see home videos about Zorro

- Can content based analysis help over structured metadata query inference?
Opportunities

Example 2:

- For main-stream head content such as news videos.
- Meta-data are fairly descriptive
- Usually queried based on non-visual attributes.

- Task: “Pull up recent Hurricane Katrina videos”
Dr. Phil McGraw visited wilt_cane Katrina's survivors...
CBSNews.com - Wednesday, September 14th, 2005
The people left devastated by Hurricane Katrina have to rebuild themselves both ph advice.
320x240 - 5min10sec - 30.0MB
www.cbsnews.com/sections/f_video/yahooVideo_frameset.html • More from this site

Hurricane Katrina is now a .... but still wreaks havoc....
CBSNews.com - Tuesday, August 30th, 2005
Hurricane Katrina has been downgraded to Category 3, but has already penetrated
Cicena, inside the Superdome and Baton Rouge.
320x240 - 4min35sec - 28.5MB
www.cbsnews.com/sections/f_video/yahooVideo_frameset.html • More from this site

Strangers welcome Hurricane Katrina's evacuees. (8/6/05)
CBSNews.com - Wednesday, September 7th, 2005
Some strangers welcome Hurricane Katrina's evacuees with open arms, reports CBS
320x240 - 3min34sec - 20.8MB
www.cbsnews.com/sections/f_video/yahooVideo_frameset.html • More from this site

Hurricane Katrina's Homeless
CBSNews.com - Friday, September 30th, 2005
One month after Hurricane Katrina, thousands of people are still left homeless by th
Chen has more on what folks in Houston shelters are facing.
320x240 - 2min49sec - 16.4MB
www.cbsnews.com/sections/f_video/yahooVideo_frameset.html • More from this site

Politicians are pointing fi lower Hurricane Katrina was...
CBSNews.com - Monday, September 19th, 2005
The blame game for the response to Hurricane Katrina has been in full swing with c
Asuras reports.
320x240 - 3min34sec - 14.9MB
www.cbsnews.com/sections/f_video/yahooVideo_frameset.html • More from this site

Former FEMA boss testified ... role in Hurricane Katrina...
CBSNews.com - Thursday, September 29th, 2005
Former FEMA Director Michael Brown testified before a House committee investigat
Katrina. Bob Orr reports.
320x240 - 2min30sec - 14.6MB
www.cbsnews.com/sections/f_video/yahooVideo_frameset.html • More from this site

Hurricane Katrina Downgraded
CBSNews.com - Tuesday, August 30th, 2005
Opportunities

Example 3:
- Creative Home Video
- Community video rendering!
- The now “famous” Star Wars Kid
- Example of “social buzz” combined with innovative tail content video production.
Opportunities

A hard example:

- Let's take an example:
  “Supposing you want to find videos that depict a monkey/chimp doing karate”!

- CBIR Approach:
  Train models for Chimps/Monkeys
  Motion Analysis for Karate movement models
  Many open issues/problems!
Video Data Management

1. Video Parsing
   - Manipulation of whole video for breakdown into key frames.
     - Scene: single dramatic event taken by a small number of related cameras.
     - Shot: A sequence taken by a single camera
     - Frame: A still image

2. Video Indexing
   - Retrieving information about the frame for indexing in a database.

3. Video Retrieval and browsing
   - Users access the db through queries or through interactions.
System overview

Query

Shot boundary detection

Feature generation

k-nn (boost, VSM)

Distances

Feature vectors

Relevance feedback

Tests data

Key frames

Retrieved results

Weighted sum of distances

ΣwD
An Architecture for Video Database System

Object Definitions (Events/Concepts)

- Spatial-Semantics of Objects (human, building, …)
- Image Features
- Object Description
- Frame

Spatio-Temporal Semantics: Formal Specification of Event/Activity/Episode for Content-Based Retrieval

- Physical Object Database
- Raw Image Database

Temporal Abstraction

- Intra/Inter-Frame Analysis (Motion Analysis)
- Inter-Object Movement (Analysis)
- Object Identification and Tracking

Sequence of Frames (indexed)

Spatial Abstraction
Video Data Management

- Metadata-based method
- Text-based method
- Audio-based method
- Content-based method
- Integrated approach
Metadata-based Method

• Video is indexed and retrieved based on structured metadata information by using a traditional DBMS

• Metadata examples are the title, author, producer, director, date, types of video.
Text-based Method

- Video is indexed and retrieved based on associated subtitles (text) using traditional IR techniques for text documents.

- Transcripts and subtitles are already exist in many types of video such as news and movies, eliminating the need for manual annotation.
Text-based Method

- Basic method is to use human annotation
- Can be done automatically where subtitles / transcriptions exist
  - BBC: 100% output subtitled by 2008
- Speech recognition for archive material
Text Detection
SHERMAN BLOCK
LOS ANGELES COUNTY SHERIFF
NCAA TEXAS TECH OKLAHOMA
8:00 ET

SHERMAN BLOCK
LOS ANGELES COUNTY SHERIFF
NCAA OKLAHOMA
8:00 ET

Final VOCR Results:
FREEMAN BLOCK
LOS ANGELES COUNT SHERIFF

SHERMAN BLOCK

SHERMAN BLOCK

SHERMAN BLOCK

SHERMAN BLOCK

O C R: S H E R M A N B L O C K

LOS ANGELES COUNTY SHERIFF

SHERMAN BLOCK

LOS ANGELES COUNTY SHERIFF

O C R: L O S A N G E L E S C O U N T S H E R I F F
Text-based Method

- Key word search based on subtitles
- Content based

Live demo:
http://km.doc.ic.ac.uk/vse/
Text-based Method

Broadcast News Browser
To view the video and subtitles, you must have the RealPlayer installed.

Click the preview frame to make it the current clip.

- Previous
- Current
- Next

Worst affected was the Calder Valley, where over 500 homes were flooded.
The River Ouse at its highest ever level for this time of year.
The river uez in York is tonight more than 14 feet above more than 14 feet above normal. People are bracing themselves.
People are bracing themselves. It’s the sort of flood expected

It's the sort of flood expected once every ten years
It's the sort of flood expected once

once every ten years in winter. But this is June - a time for lazy
But this is June - a time for lazy days and holidays, which days and holidays, which for thousands has turned into a

It’s the sort of flood expected once every ten years
outline

• images
• video

• speech
Spoken Document Retrieval

**Acoustic Modeling**
Describes the sounds that make up speech

**Speech Recognition**

**Lexicon**
Describes which sequences of speech sounds make up valid words

**Language Model**
Describes the likelihood of various sequences of words being spoken
Speech Recognition in Brief

Signal Processing → Phonetic Probability Estimator (Acoustic Model) → Decoder (Language Model) → Grammar

Pronunciation Lexicon
Hints For Better Recognition

• Goal: improve the estimation $p(\text{word}|\text{acoustic\_sig})$
• Main idea:
  
  $p(\text{word}|\text{acoustic\_sign}) \rightarrow p(\text{word}|\text{acoustic\_signal, X})$

  What could be $X$?

• Topical information
• News of the day
• Image information?
Hints For Better Recognition

• Goal: improve the estimation $p(\text{word}|\text{acoustic\_sig})$
• Main idea:
  $p(\text{word}|\text{acoustic\_sign}) \rightarrow p(\text{word}|\text{acoustic\_signal, X})$

What could be $X$?

• Topical information
• News of the day
• Image information
  – Lip reading
  – Video Optical Character Recognition (VOCR)
Information Retrieval Precision vs. Speech Accuracy

A rather small degradation in retrieval when word error rate is small than 30%

Indexing and Search of Multimodal Information, Hauptmann, A., Wactlar, H.
Spoken Document Retrieval: Document Expansion

• Motivation: documents are erroneous (or ambiguous)
• Goal: apply expansion techniques to correct the word errors in documents
• Similar to query expansion
  – Treat each speech document as a query
  – Find clear documents that are relevant to speech documents
  – Expand each speech document with the words that are common in the clear documents that are relevant.
Demos

http://images.google.com/

http://video.google.com/

http://www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English

http://amazon.ece.utexas.edu/~qasim/research.htm

http://mp7.watson.ibm.com/

http://viper.unige.ch/research/video/
END
exam topics

• Evaluation
  – Recall, precision, E, F
  – AP
  – R-prec
  – PCutoff
  – precision-recall curves
• Retrieval models
  – Boolean
  – Vector space
  – Language modeling
  – Inference networks
• Indexing
  – Manual vs. automatic
  – Tokens, stopping, stemming,
• File organization
  – Bitmaps
  – Signature files
  – Inverted files
• Statistics of text
  – Zipf's law
  – Heap's Law
  – Information theory

• Compression
  – Huffman
  – LempelZiv
• Relevance feedback
  – Real
  – Assumed
• Clustering
  – Graph, partitioning, nearest neighbor
  – clustering algorithms
• Markov chains
  – stationary distribution
• Page Rank formula
• Metasearch
  – CombSUM
  – Borda
  – Condorcet
• Collaborative filtering
• P2P
  – 3 generations