Multimedia semantic computing: content analysis, recognition and authoring

Yu Hen HU
University of Wisconsin - Madison
Dept. Electrical and Computer Engineering
1415 Engineering Drive
Madison, WI 53706
hu@engr.wisc.edu
Outline

• What is multimedia Content Analysis and Recognition
• State of Art
• Future Direction
What is Multimedia Content?

• **Content**
  - (in the context of media and publishing) information and experiences created to benefit audiences in contexts that they value - Wikipedia

• **Digital media object**
  - Article, song, picture, movie, etc. represented in digital formats

• **Content of digital media objects**
  - Description of low level features (syntax)
  - Concepts and Context (semantics)
Low Level Features

• Described in MPEG-7

• Examples
  - Images:
    • Color: color space, dominant color,
    • Texture
    • Shape
  - Video (see figure)
  - Audio
    • Silence, sound effects, spoken language, timbre, melody contour

(C) 2008 by Yu Hen Hu
High Level Features

• Concept, objects, context
• Difficult for generic classes, but possible for domain specific recognition
• Example concepts (TrekVid) from TV News
  - Sports, weather, office, meeting, desert, mountain, waterscape/waterfront, corporate leader, police security, military personnel, animal, computer tv screen, us flag, airplane, car, truck, people marching, explosion fire, maps, charts
  - Part of LSCOM (Large scale ontology for multimedia)
• Domain specific examples:
  - Specific sports: home run, strike, etc.
Multimedia Content Processing

• Summary,
  - Generation of multimedia program guide or content summary
  - Generation of content description of A/V archive to allow seamless exchange among content creator, aggregator, and consumer.
• Filtering
  - Filter and transform multimedia streams in resource limited environment by matching user preference, available resource and content description.
• Retrieval
  - Recall music using samples of tunes
  - Recall pictures using sketches of shape, color movement, description of scenario
• Recommendation
  - Recommend program materials by matching user preference (profile) to program content
• Indexing
  - Create family photo or video library index

(C) 2008 by Yu Hen Hu
MM Content Processing Efforts

- **Content Based Image Retrieval**
  - Retrieve desired pictures using low level features
  - Considered matured technology except HCI is still unnatural
  - Potential applications
    - Finding product of identical patterns, similar shape

- **Content based song/music retrieval**
  - Retrieve desired songs using a single phrase of tune
  - Some success
  - Applications: unclear (KTV?)

- **Content based video retrieval**
  - Use audio/speech cue and other video feature to retrieve videos
  - Applications on news
  - TREKVID competition

(C) 2008 by Yu Hen Hu
MM Content Search Applications

• Content based image retrieval (CBIR)
  - Organize amateur photo library
  - Looking/shopping for objects of specific color, texture or shape
    • Matching set of dining set, matching color of paints, matching carpet, drapery, etc.
  - Identify unknown object/building/person in a photo
    • By finding labeled picture with similar/identical object
• Multimedia authoring
  • Background picture, inserts that are contextually relevant to current text
MM Content Search Applications

- Content based audio retrieval
  - Organizing amateur audio library
    - E.g. IPOD library
  - Disk jockeying, music program recommendation
  - Find a song/music piece based on partial melody, lyric
  - Identify unknown sound source
    - Forensic scenario,
  - Video background sound effect generation
    - Sound of specific event/action
    - Sound suitable for the mood/context/tempo of scene
**MM Content Search Application**

- **Content based video retrieval**
  - Amateur video library organization
    - Home video
    - Personal movie collection
    - YouTube collection search
  - Video search engines
    - Yahoo, Google, etc.
  - Mobile video over cell phone, PDA, and other appliance
  - Surveillance video search (security)
- **Authoring**
  - Search for relevant video for a TV news story
  - Creating a new video based on relevant video clips
    - Such as those in discovery channel/national geographics
A Framework of Generic Content-based Retrieval

Query Module
- Feature extraction
- Interactive Query Formation

Retrieval Module
- Feature comparison
- Browsing & Feedback

Input Module
- Feature extraction
- Multimedia data

User ➔ Feature Database ➔ Image Database ➔ Output

(C) 2008 by Yu Hen Hu
Content-Based Visual Query (1)

• Advantage
  - Ease of creating, capturing and collecting digital imaginary

• Approaches
  - Extract significant features (Color, Texture, Shape, Structure)
  - Organize Feature Vectors
  - Compute the closeness of the feature vectors
  - Retrieve matched or most similar images
Content-Based Visual Query (2)  
Improve Efficiency

• Keyword-based search
  - Match images with particular subjects and narrow down the search scope

• Clustering
  - Classify images into various categories based on their contents

• Indexing
  - Applied to the image feature vectors to support efficient access to the database
State of Art of Image Retrieval

• Content based image retrieval tools available
  - Using color, sketch
  - Using image examples
    • Whole image
    • Patch of image
  - Keyword (concept) based search

• Limitations of current tools
  - Ability to segment individual object in an image can still be improved
  - Recognition of particular object (e.g. face) depends on quality of image (resolution, color, occlusion, etc.)
Google Image Search: Typhoon

(C) 2008 by Yu Hen Hu
**ImgSeek**

- imgSeek is a photo collection manager and viewer with content-based search and many other features.
- The query is expressed either as a rough sketch painted by the user or as another image you supply (or an image in your collection).
- The searching algorithm makes use of multiresolution wavelet decomposition of the query and database images.
ImgSeek: Query by Color
ImgSeek: Query by Sketch
ImgSeek: Query by Example
ImgSeek: Organize Family Album

File Collection Help
Browse | Add | Search | Options |
by Image content | by Keyword |

Field | Value
1 | Description | cat
2 | Description | dog

/home/rc/doc/mavica-fotos/others/DSC0003
4.JPG
/home/rc/doc/mavica-fotos/others/DSC0003
3.JPG
/home/rc/doc/mavica-fotos/others/DSC0003
5.JPG
/home/rc/doc/mavica-fotos/others/DSC0003
6.JPG
/home/rc/doc/mavica-fotos/others/DSC0003
6.JPG
/home/rc/doc/mavica-fotos/others/DSC0003
7.JPG

104 images | /home/rc/doc/mavica-fotos/others/DSC0003.JPG

(C) 2008 by Yu Hen Hu
Content Based Video Retrieval

- Many on-line video libraries exist
- Use primarily text based search (tags)
- Non use content based features
  - Difficulty in UI design
  - Both visual and audio cues
- TrekVid contest has gain much attention
- New direction
  - automatically extract keywords (concept) from video clips (need vocabulary)
  - Retrieval using hierarchy of concept (context) for retrieval
    - Location

(C) 2008 by Yu Hen Hu
Example: Yahoo Video

(C) 2008 by Yu Hen Hu
Yahoo Video Advance Search
Google Video Advanced Search Interface
Blinkx Advanced Search Options
YouTube
Segmentation

Given observation $x_k$, estimate probability $p(\text{story bnd} = \text{YES} \mid x_k)$

- anchor face?
- visual motion?
- video caption text?
- music or speech?
- new speech segment?
- significant pause?
- pitch change?
- cue phrases appear?
Use Low Level Feature Find Similar Images

• Given a query image
• Extract low level features
  • Color
  • Texture
  • Edge
• Evaluate “distance” in the feature space between query pictures and those of gallery pictures
• Top ranked matches (closest ones) are displayed for user selection.
• User selection becomes “relevance feedback” to improve future search accuracy

(C) 2008 by Yu Hen Hu
Vision based Approach: SIFT

- Scale Invariant Feature Extraction Method
- Identify (automatically) outstanding features
  - Sharp edges, low straight lines, etc.
  - Mostly geometrically invariant features (differential invariants)
- Perform pair-wise feature matching using RANSAC
  - A random search method to address the NP hard issue of feature matching
Feature detection

Detect features using SIFT [Lowe, IJCV 2004]
Feature detection

Detect features using SIFT [Lowe, IJCV 2004]
Feature detection

Detect features using SIFT [Lowe, IJCV 2004]
Feature matching

Match features between each pair of images
Feature matching

Correspondence estimation

- Link up pairwise matches to form connected components of matches across several images
TREKVID

• TREC Video Retrieval Evaluation
  - Grow out of Text Retrieval Conference (TREC)
  - Focus on video retrieval
  - Organized by US NIST

• Objectives
  - to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results

• Started in 2001 as special track of TREC.
• Become independent workshop since 2003.

http://www-nlpir.nist.gov/projects/trecvid/
TREKVID 2006 Statistics

• Data
  - 159 hrs (Nov/Dce.’05 news in Arabic, Chinese, English)
  - 50 hrs of BBC rushes (unedited video footages)

• On news data, 3 evaluation tasks
  - Shot boundary determination
  - High level feature extraction (39 features submitted, 20 evaluated)
  - Search (automatic, manually-assisted, interactive)
    - Base scenario: an English-only searcher looking through videos in Arabic, Chinese, and/or English

• 1 exploration tasks on BBC rushes
  - Identify and remove redundancy
  - Organize/present according to useful features
  - Devise a practical informative evaluation scheme

(C) 2008 by Yu Hen Hu
Shot Boundary Detection Test

- to identify the shot boundaries with their location and type (cut or gradual) in the given video clip(s)

- Data
  - 13 representative news videos, 597043 frames, 3785 transitions:
    - 1,844 (48.7%) Cuts (2005: 60.8%)
    - 1,509 (39.9%) Dissolves (2005: 30.5%)
    - 51 (1.3%) Fadeout/in (2005: 1.8%)
    - 381 (10.1%) other (2005: 6.9%)
    - More graduals, which are harder to match
Shot Boundary Test Results
(cut, top performers)
Gradual Transition Detection
(top Performers, zoomed)
Run Time Comparison

![Run Time Comparison Graph]

(C) 2008 by Yu Hen Hu
High Level Feature Results

category A, upper half

TRECVID 2006

(C) 2008 by Yu Hen Hu
Inferred Average Precision by Features

Feature number

Average precision

1 sports 26 animal
3 weather 27 computer tv screen
5 office 28 us flag
6 meeting 29 airplane
10 desert 30 car
12 mountain 32 truck
17 waterscape/waterfront 35 people marching
22 corporate leader 36 explosion fire
23 police security 38 maps
24 military personnel 39 charts

(C) 2008 by Yu Hen Hu
TREKVID Search Tasks

AUTOMATIC:

System takes topic as input and produces result without any human intervention.

MANUAL:

Human formulates query based on topic and query interface, not on knowledge of collection or search results.

INTERACTIVE:

Human (re)formulates query based on topic, query, and/or results.

System takes query as input and produces result without further human intervention.

Number of runs:
- 76 automatic
- 11 manually assisted
- 36 interactive
181. Find shots of one or more soldiers or police with one or more weapons and military vehicles [2, 6, 128]
182. Find shots of water with one or more boats or ships [3, 5, 307]
183. Find shots with one or more emergency vehicles in motion (e.g., ambulance, police car, fire truck, etc.) [0, 4, 299]
184. Find shots of one or more people seated at a computer with display visible [3, 4, 440]
185. Find shots of one or more people reading a newspaper [3, 4, 201]
186. Find shots of a natural scene with, for example, fields, trees, sky, lake, mountain, rocks, rivers, beach, ocean, grass, sunset, waterfall, animals, or people; but no buildings, no roads, no vehicles [2, 4, 523]
187. Find shots of one or more helicopters in flight [0, 6, 119]

[ number of image, video examples and relevant found]
Example Search Results

Unique relevant shots return by Oxford U. for Topic 191 ("adult and child")

(C) 2008 by Yu Hen Hu
2006: Average precision by topic

![Graph showing average precision by topic over time, with events marked on the x-axis and mean average precision on the y-axis. The graph includes lines for different methods such as interactive max, manual max, and automatic max, with median values also shown.](image)

(C) 2008 by Yu Hen Hu
Future Directions of Image/Video Search

- Current image database rely on text captions
  - Keywords may be context dependent
    - Eg. Typhoon search in google news
  - Community based annotation
    - Web 2.0 concept, eg. Wikipedia

- Query by example
  - Provide an example or portion of an image, find all that similar or contain the same portion
  - Sometimes useful but limited applications
    - Matching patterns/colors

- Concept learning
  - Exploit generic concept automatically to create keywords
  - Manually add context dependent keywords
  - Text based search

(C) 2008 by Yu Hen Hu
Conclusion

• Content based multimedia retrieval
  - Technology:
    • Content based search seems feasible, but cannot be used alone. Need to complemented with text based meta information
    • Holy grail problem:
      - Inferring context (meta text) from raw data
    • Possible solution
      - Automatic extraction + user annotation
  - Business model:
    • Scenario abound, but still need to be explored.