State-of-the-art Multimedia Search Engines

- Work better for simple concepts,
  e.g. Two people kissing, A picture of a giraffe
- Don’t work for complex queries
  e.g. A picture of a brick home with black shutters and white pillars, with a pickup truck in front of it (image)
Examples

- Find the pictures of giraffe
  - Keyword: giraffe

- A picture of a brick home with black shutters and white pillars, with a pickup truck in front of it (image)
  - brick home shutters
In Google Video try searching for “Bush” or “Bush speaking about Iraq”
Why this happens?

- Most of these search engines are keyword based
  - "False" multi-media search engine
  - Have to represent your idea in keywords
  - These keywords are expected to appear in the filename, or corresponding webpage

- Therefore......
  - Unable to handle semantic meaning of images
  - Unable to handle visual position
  - Unable to handle time information
  - Unable to use images as query
  - ............
"When I use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean—neither more nor less.

Try to search the image of the logo of Osama
How Google does it?

- No image processing. Textual context!
  - In videos it uses closed captions and transcriptions
- File names, nearby words
- Distance from image to words
- “give me images with flower in the file name or near the image”
Solution

- "...it would be great to have multimedia search engine intelligent enough to associate its own keywords based on what's in the image."

- "Content-based information retrieval (CBIR)"
Different from text IR:
- Structure of data is more complex. Efficiency is an issue
- Using of metadata
- Characteristics of multimedia data
- Operations to be performed

Aspects:
- Data modeling: Extract and maintain the features of objects
- Data retrieval: based not only on description but on content
Retrieval process

- Query specification
  - fuzzy predicates: similar to
  - content predicates: images containing an apple
  - data type predicates: video, ...

- Query processing and optimization
  - Parsed, compiled, optimized for order of execution
  - Problem: many data types, different processing for each

- Answer
  - Relevance: similarity to query

- Iteration
  - Bad quality, so need to refine
Multimedia Information Retrieval is quite big in scope:

Data examples:
- 2D/3D color/grayscale images: e.g., brain scans, scientific databases of vector fields
- (2D) video,
- (1D) voice/music; (1D) time series: e.g., financial/marketing time series; DNA/genomic databases

Query examples:
- find photographs with the same color distribution as this
- find companies whose stock prices move as this one
- find brain scans with a texture of a tumor
- Find videos where something happens
Some solutions

- Reduce the problem to search for multi-dimensional points (feature vectors, but vector space is not used)
- Define a distance measure
  - for time series: e.g., Euclidean distance between vectors
  - for images: e.g., color distribution (Euclidean distance); another approach: *mathematical morphology*
  - Other features as vectors
- Often, for search within distance, the vectors are organized in R(*/+)/trees or other spatial trees
- Clustering plays important role
Query types

- All within given distance
  - Find all images that are within 0.05 distance from this one

- Nearest-neighbor
  - Find 5 stocks most similar to IBM

- All pairs within given distance
  - Further: clustering

- Whole object vs. sub-pattern match
  - Find parts of image that are...
  - E.g., in 512 × 512 brain scans, find pieces similar to the given 16 × 16 typical X-ray of a tumor
  - Like passage retrieval for text documents
Open problems

- How similarity function can be defined?
- What features of images (video, sound) there are?
- How to better specify the importance of individual features? (*Give me similar houses*: similar = size? color? structure? Architectural style?)
- How to determine the objects in an image?
- Integration with DBMSs and SQL for fast access and rich semantics
  - Integration with XML
  - Ranking: by similarity, taking into account history, profile
Open problems

- Object/event detection (computer vision and pattern recognition)
- Automatic feature selection
- Spatial indexing data structures (more than 1D)
- New types of data.
  - What features to select? How to determine them?
- Mixed-type data (e.g., web pages, or images with sound and description)
- What clustering/IR methods are better suited for what features? (What features for what methods?)
- Similar methods in data mining, ...
Content-based Video Retrieval

- Application
- Implementation
- Experience from TREC video track
  - Feature Extraction Task (High-level Semantics Feature)
  - Manual Retrieval Task (One-run Retrieval)
  - Interactive Retrieval Task (Multiple-run with Feedback)
  - Results & Demo (CMU and IBM)
- Conclusion
Application

- Increasing demand for visual information retrieval
  - Retrieve useful information from databases
  - Sharing and distributing video data through internet

Example: BBC
- BBC archive has +500k queries plus 1M new items ... per year;
- From the BBC ...
  - Police car with blue light flashing
  - Government plan to improve reading standards
  - Two shot of Kenneth Clarke and William Hague
Application ( Cont. )

- Past project: ASSAVID in collaboration with BBC sports library:
  - Develop automatic annotation systems for sports videos
Application ( Cont. )

- Video Surveillance
  - Find where else the person appears
- Experience On-Demand
  - Help to remember previous events
- Provide useful information on traveling
  - Equipment on cars to retrieve useful multimedia information according to your location/preference

- ...........

- Video content is plentiful ... its now available digitally ... we can work on it directly ... so it follows
Application (Cont.)

Featured Videos

The Tonight Show w/ Jay Leno -- Phony Photo Booth
04:51
The Tonight Show w/ Jay Leno puts a phony photo booth in at Universal Studios in Hollywood. It starts off slow but gets pretty damn funny! Catch Jay every weeknigh at 11:35pm only on NBC.
Tags: Tonight Show Jay Leno Phony Photo Booth funny comedy
Added: 1 week ago in Category: Entertainment
From: NBC
Views: 6790
541 ratings

Mr. Pringles Press Conference
02:28
Mr. Pringles makes a public apology for his recent arrest
Tags: mr. pringles gail gibson public apology press conference michael booth fadem assenators wpea
Added: 6 days ago in Category: Comedy
From: wpea
Views: 137775
1516 ratings

Biff's Question Song (Stand-Up Comedy)
01:52
Have a question about Back to the Future for the guy who played "Biff"? Yeah, so do a billion other people! Here's a song ... my little time saver! www.TomWilsonUSA.com

Cingular Presents: YouTube Underground - Submissions Open
Login to rate video
From: underground
Comments: 39
365 ratings

Member Login
User Name: 
Password: 
Login
Sign Up
Application (Cont.)
Typical Retrieval Framework

- **User**: provide query information that represents his information needs
- **Database**: store a large collection of video data
- **Goal**: Find the most relevant shots from the database
- **Shots**: “paragraph” in video, typically 20 – 40 seconds, which is the basic unit of video retrieval
Bridging the Gap

Video Database

User

Result
Automatically Structure Video Data

- The first step for video retrieval: Video “programmes” are structured into logical scenes, and physical shots.
- If dealing with text, then the structure is obvious:
  - paragraph, section, topic, page, etc.
- All text-based indexing, retrieval, linking, etc. builds upon this structure;
- Automatic shot boundary detection and selection of representative keyframes is usually the first step;
Typical automatic structuring of video

A video document

A set of shots

Keyframe browser combined with transcript or object-based search
Bridging the Gap

Video Database

Video Structure

User

Information Need

Result
Ideal solution

Video Database

Video Structure

Understanding the semantic meaning and retrieve

Result

User

Information Need
Ideal solution

Video Database

Video Structure

Understanding semantic meaning and retrieve

However,
1. Hard to represent query in natural language and for computer to understand
2. Computers have no experience
3. Other representation restriction like position, time

Result
Alternative Solution

Video Database

Video Structure

Provide evidence of relevant information (text, image, audio)

Match and combine

Result

User

Information Need
Evidence-based Retrieval System

- General framework for current video retrieval system
- Video retrieval based on the evidence from both users and database, including
  - Text information
  - Image information
  - Motion information
  - Audio information
- Return a relevant score for each evidence
- Combination of the scores
Keyword-based System

Video Database

Video Structure

Automatic Annotation

Including filename, video title, caption, related web page

Keyword

User

Information Need
Keyword-based System

Video Database

Video Structure

Automatic Annotation

Manual Annotation

Keyword

User

Information Need
Manual Annotation

- Manually creating annotation/keywords for image / video data
- Examples: Gettyimages.com (image retrieval)
- Pros:
  - Represent the semantic meaning of video
- Cons
  - Time-consuming, labor-intensive
  - Keyword is not enough to represent information need
Manual annotation using metadata

Try the open source version:
lulop2.sf.net
Tagging

Hello, world! This is a channel of Lulop.com built with the sole purpose of demonstrating the concept of Shot Tagging, which is indexing individual scenes and sections within a video, in this particular case using a video player. Use tags on the left to select your video. Play the video you have selected and the relevant scene will be added to the player as it is.

Tags: lulop.com purely a demo page.

“The Color Purple” afterparty Hot Girls

Lohan Moms Day Nicky’s Moms lunch

Valentino Marc! Venice Cinema Festival Venice 2008

“Detained” premiere Lindsay Lohan is there

“Match Point” photocall “Match Point” Press Conference

“Un posta al sole,” 160 F 44th St, New York City, 07/02/08 Pregnancy 07/02 50th Birthday of King Carl XVI Gustaf

Film Festival Aaron Aaron Carter Acadamy Awards

acapulco ADAM YOUNG ad Adam Brody ad

Adrian Brody ADRIAN GRENIER ad

Restaurant aircraft Airport Alan Fawcett

WHATEVER YOU BRING WE SING Alcohol

Alessandro Fernandez Alessandro Bottone

Alessandro Zanni ALISON MELNICK Allis

MOLLASSO Anastasia Andre Heller And

Andy Roddick Andy Lee Angela Bassett
Manual annotation using taxonomy (sort of...)
Manual annotation using taxonomy

Taxonomy is exploited for retrieval
Manual annotation using 4 Ws

Who

What

Where

When

Frameline uses MPEG-47: i.e. MPEG-4 with MPEG-7 stream
Speech and OCR transcription

Video Database

Video Structure

Annotation

Speech Transcription

OCR Transcription

Keyword

Information Need

User
Query using speech/OCR information

**Query:**
Find pictures of Harry Hertz, Director of the National Quality Program, NIST

**Speech:**
We’re looking for people that have a broad range of expertise that have business knowledge that have knowledge on quality management on quality improvement and in particular …

**OCR:**
Harry Hertz a Director aro 7 wa-, i,, ty Program ,Harry Hertz a Director
Automatic face and OCR recognition

Riya's search looks inside of photos to recognize:

1. People

![Image of a person with name tag]

2. Text

![Image of a sign with text]

**People**

- Deven

**Text**

- Tomichi Point
Closed captions

View selected videos with captions

**Blue Oyster Spiral Fractal Zoom**
No matter how deep you go, there's always more. This movie took quite a few days to calculate. In 1993, when I first made a poster of the image I ...  
Dave Kliman - 3 min - Jun 9, 2006  
[3 stars] (284 ratings)

**Biofuels: Think Outside The Barrel**
Google TechTalks March 29, 2006 Vinod Khosla Vinod Khosla is a venture capitalist considered one of the most successful and influential personalities ...  
Google engEDU - 1 hr 9 min - Mar 29, 2006  
[3 stars] (249 ratings)

**NOVA: This Old Pyramid**
NOVA reveals the secrets of how the ancient pyramids were built by actually building one. A noted Egyptologist, Mark Lehner, and a professional stonemason, ...  
WGBH Educational Foundation - 56 min - Nov 4, 1992  
[3 stars] (70 ratings)
DVD subtitle ripping and OCR

Deja correr bastante el agua.
Podría obstruirse.
Lo siento.
What we lack?

Video Database

Video Structure

Annotation

Speech Transcription

OCR Transcription

Keyword

Image Information

User

Information Need
Image-based Retrieval

Video Database

User

Video Structure

Text Information

Keyword

Query Images

Information Need

Image Feature
Global Low-level Image Feature

- Color-based Feature
  - Color Histogram
  - Color Percentage
  - Color Correlogram
  - Color Moments
- Texture-based Feature
  - Gabor Filter
  - Wavelet
- Shape/Structure Feature
Regional Low-level Image Feature

- Segmentation into objects – hard problem!
- Extract low-level features from each region
Image Search

- Feature Representation
  - Image: represented as a series of real number, or a vector of features, \((f_1, \ldots, f_n)\)
  - Distance Function: The distance between two vectors, typically Euclidean Distance
  - Probably “Nearest is relevant”
    - The nearest images in the database is relevant to the query images.
Finding Similar Images
But.....

- Low-level feature doesn’t work in all the cases
Find similar objects
High-level Image Feature

- Objects: Persons, Roads, Cars, Skies…
- Scenes: Indoors, Outdoors, Cityscape, Landscape, Water, Office, Factory…
- Event: Parade, Explosion, Picnic, Playing Soccer…
- Generated from low-level features
Image-based Retrieval

Video Database

Video Structure

Text Information

Image Feature

Low-level Feature

High-level Feature

Keyword

Query Images

User

Information Need
More Evidence in Video Retrieval

Video Database

Video Structure

Text Information

Image Information

Motion Information

Audio Information

Keyword

Query Images

Motion

Audio

User

Information Need
Combination of multi-modal results

- Difference characteristics between multi-modal information
  - Text-based Information: better for middle and high level queries
    - e.g. Find the video clip of dancing women wearing dresses
  - Image-based Information: better for low and middle level queries
    - e.g. Find the video clip of green trees

- Combination of multi-modal information
Other Useful Technique

- Query Expansion
- Cross-Modal Relation
- Relevance Feedback
Recap

- Video Retrieval is to bridge the gap between user information need and video database
- Multi-modal evidence
  - Text-based (most popular)
  - Image-based
  - Motion-based
  - Audio-based
- Combination of the evidence
Introduction to TREC Video Retrieval Track

- Full Name: Text REtrieval Conference
- TREC Video Track web site: http://www-nlpir.nist.gov/projects/trecvid/
- TREC series sponsored by the National Institute of Standards and Technology (NIST) with additional support from other U.S. government agencies
  - Goal is to encourage research in information retrieval
Introduction to TREC Video Retrieval Track

- Video Retrieval Track started in 2001
  - Goal is investigation of content-based retrieval from digital video
  - Focus on the shot as the unit of information retrieval rather than the scene or story/segment/clip

- Current state-of-the-art Video Retrieval Competition
  - 17 active participants, including groups from CMU, IBM Research, Microsoft Research Asia, MediaMill, LIMSI, Dublin City University.
Main tasks in TREC

- Shot boundary detection
- Semantic Feature Extraction Task
- Video Retrieval Task
  - Manual Retrieval: Human formulate a query and then automatically retrieve from collection
  - Interactive Retrieval: Full human access and feedback
Where are they?

Video Database

Video Structure

Shot Boundary Detection

Retrieval Task

Keyword

Information Need

User

Query Images

Feature Extraction

Text Information

Image Feature

Low-level Feature

High-level Feature
Video Data

- Difficult to get video data for use in TREC because ©
- Used mainly Internet Archive
  - advertising, educational, industrial, amateur films 1930-1970
  - produced by corporations, non-profit organisations, trade groups, etc.
  - Noisy, strange color, but real archive data
  - 73.3 hours partitioned as follows:
Shot Boundary Detection

- Fundamental primitive of most/all work in content-based video retrieval
Feature Extraction

- Extracted high-level semantic feature from video
- Assign a video clip to one or more of several categories of video

High-level features: Cityscape, Lake, Trees, Water, Sky
Feature Extraction

• Interesting itself but when it serves to help video navigation and search then its importance increases

• Benefits:
  
  • Retrieval - Find video from a particular class
  
  • Filtering - Remove irrelevant and distracting information categories from summaries and visualizations
The Features

- **Face**
  - Clip contains at least one human face with the nose, mouth, and both eyes visible. Pictures of a face meeting the above conditions count.

- **People**
  - Clip contains a group of two more humans, each of which is at least partially visible and is recognizable as a human.

- **On-screen Text**
  - Clip contains superimposed text large enough to be read.
The Features

Indoor
Clip contains a recognizably indoor location, i.e., inside a building

Outdoor
Clip contains a recognizably outdoor location, i.e., one outside of buildings

Cityscape
Clip contains a recognizably city/urban/suburban setting

Landscape
Clip contains a predominantly natural inland setting, i.e., one with little or no evidence of development by humans. Scenes with bodies of water that are clearly inland may be included
Non-Video (Audio) Features

Speech
A human voice uttering words is recognizable as such in this segment

Instrumental Sound
Sound produced by one or more musical instruments is recognizable as such in this segment

Monologues
Segment contains an event in which a single person is at least partially visible and speaks for a long time without interruption by another speaker. Pauses are ok if short
TREC02 Results

The diagram shows the average precision for different features across various categories. The categories include:

- Outdoors
- Indoors
- Face
- People
- Cityscape
- Landscape
- Text overlay
- Speech
- Instrumental sound
- Monolog

The features are represented on the x-axis and the average precision on the y-axis. The lines and symbols indicate the performance of different systems, such as CMU_r1, A_CMU_r2, CLIPS-LIT_GEOD, CLIPS-LIT-LIMSU, etc. The random baseline is also indicated by a dotted line at the bottom of the graph.
Video Search Task

- The most important task and final goal
- Manual & Interactive Search Task
Queries for 2002 TREC Video Track

- **Specific item or person**
  - Eddie Rickenbacker, James Chandler, George Washington, Golden Gate Bridge, Price Tower in Bartlesville, OK

- **Specific fact**
  - Arch in Washington Square Park in NYC, map of continental US

- **Instances of a category**
  - football players, overhead views of cities, one or more women standing in long dresses

- **Instances of events/activities**
  - people spending leisure time at the beach, one or more musicians with audible music, crowd walking in an urban environment, locomotive approaching the viewer
TRECVid 2005 search topics

TRECVID 2005 Interactive Video Retrieval Results

1. Condoleezza Rice
2. Iyad Allawi
3. Omar Karmi
4. Hu Jintao
5. Tony Blair
6. Mahmoud Abbas
7. Graphic map of Iraq, Baghdad marked
8. Two visible tennis players on the court
9. People shaking hands
10. Helicopter in flight
11. George W. Bush entering or leaving a vehicle
12. Something on fire with flames and smoke
13. People with banners or signs
14. People entering or leaving a building
15. A meeting with a large table and people
16. A ship or boat
17. Basketball players on the court
18. One or more palm trees
19. An airplane taking off
20. A road with one or more cars
21. One or more military vehicles
22. A tall building
23. A goal being made in a soccer match
24. Office setting

48 users of other video retrieval systems

Proposed lexicon-driven MediaMill system
TRECvid 2006

FULLY AUTOMATIC:

**TOPIC** → **SYSTEM** → **RESULT**

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

MANUALLY-ASSISTED:

**TOPIC** → **HUMAN** → **QUERY** → **SYSTEM** → **RESULT**

Human formulates query based on topic and query interface, not on knowledge of collection or search results. System takes query as input and produces result without further human intervention.

INTERACTIVE:

**TOPIC** → **HUMAN** → **QUERY** → **SYSTEM** → **RESULT**

Human (re)formulates query based on topic, query, and/or results. System takes query as input and produces result without further human intervention on this invocation.
Queries for 2006 TREC Video Track

Example types of video needs

I'm interested in video material containing:

- a specific person
- one or more instances of a category of people
- a specific thing
- one or more instances of a category of things
- a specific event/activity
- one or more instances of a category of events/activities
- a specific location
- one or more instances of a category of locations
- combinations of the above

Topics may target commercials as well as news content.
Some TRECVid 2006 high level features

- **Sports**: Shots depicting any sport in action
- **Entertainment**: Shots depicting any entertainment segment in action
- **Weather**: Shots depicting any weather related news or bulletin
- **Court**: Shots of the interior of a court-room location
- **Office**: Shots of the interior of an office setting
- **Meeting**: Shots of a Meeting taking place indoors
- **Studio**: Shots of the studio setting including anchors, interviews and all events that happen in a news room
- **Outdoor**: Shots of Outdoor locations
- **Building**: Shots of an exterior of a building
- **Desert**: Shots with the desert in the background
- **Vegetation**: Shots depicting natural or artificial greenery, vegetation woods, etc.
- **Mountain**: Shots depicting a mountain or mountain range with the slopes visible
- **Road**: Shots depicting a road
- ...
Sample Query

- XML Representation

```xml
<!DOCTYPE videoTopic SYSTEM "videoTopics.dtd">
<videoTopic num="077">
  <textDescription text="Find pictures of George Washington" />
  <imageExample src="http://www.cia.gov/csi/monograph/firstLn/955pres2.gif" desc="face" />
  <videoExample src="01681.mpg" start="09m25.938s" stop="09m29.308s" desc="face" />
</videoTopic>
```
Evaluation Metric

Goal: Maximize the Mean Average Precision

- Result set limited to 100 shots
- Precision = (# relevant shots retrieved)/(total # shots retrieved)
- Average precision: compute precision after each retrieved relevant shot and then average these precisions over the total number of retrieved relevant shots in the collection for that topic
- Submitting the maximum number of shots per result set can never lower the average precision for that submission
- Mean Average Precision = average of the average precision measures for each topic
Demo

- CMU Interactive Search System
- IBM Video Retrieval System
- UvA MediaMill
CMU Manual Retrieval System

Query

Text
  - Movie Info
  - Text Score

Image
  - Image Score
  - PRF Score

Final Score

Retrieval Agents
Snapshot of the CMU system (2002)
Snapshot of the CMU system (2002)
CMU Filter Interface for using Image Features (2002)
Mean AvgP vs. mean elapsed time

Wide variation in elapsed time. Not the dominant factor in effectiveness

Mean average precision

Mean elapsed time (mins.)
IBM Marvel system

Search using color similarity
IBM Marvel: search using semantics from previous results
UvA MediaMill

TRECVID 2004 Interactive Search Results

Mean Average Precision

System Runs

32 concept detectors

Interactive video search

TRECVID 2005 Interactive Search Results

Mean Average Precision

System Runs

101 concept detectors
UvA MediaMill – cross browser
Conclusion

- The goal of content-based video retrieval is to build more intelligent video retrieval engine via semantic meaning.
- Many applications in daily life.
- Combine evidence from different aspects.
- Hot research topic, few business system
  - Check Techcrunch.com for info on business ventures.
- State-of-the-art performance is still unacceptable for normal users, space to improve.
Virage: a business that has survived
Reply.it – Multimedia asset management

MEDIA INGESTION
- Encoding
- Indexing
- Quality Control
- Content Provider Adaption
- Legacy System Integration

DIGITAL ASSET MANAGEMENT
- Logging
- Searching
- Transcoding
- Archiving (HSM)
- Browsing
- Rich Media Editing
- Cut Edit

DISTRIBUTION
- App Services
- Web Services
- Messaging Services
- Rendering Services

AIR DISTRIBUTION
- Integration
- Witness

Internet: (web)
Mobile: (Handheld/3G/...)
Broadcast: (TV/ iTV/DVB-T)
Credits

- Rong Yan – Carnegie Mellon
- Alexander Gelbukh