Multimedia Information Retrieval
The case of video
Outline

- Overview
- Problems
- Solutions
- Trends and Directions
Motivation

- With the explosive growth of digital media data, there is a huge demand for new tools and systems that enables average users to more efficiently and more effectively search, access, process, manage, author and share these digital media contents.
Multimedia Information Retrieval

Text-based Information Retrieval
- Too many images to annotate
- High cost of human interpretation
- Subjectivity of visual content, e.g., “A picture is worth a thousand words”

Content-based Retrieval
- automatically retrieves images, video, and audio based on the visual and audio content

History
- Conference on Database Applications of Pictorial Applications in 1979
- NSF workshop in 1992
- More active field since 1997 when Internet and web browsing became popular
A VERY DIVERSIFIED FIELD!

- **Data Types**
  - Text, hypertext, image, audio, graphics, animation, paintings, video/movie, rich text, spread sheet, slides, combinations of these and user interaction

- **Research Problems**
  - Systems, content, services, user, evaluation, implementation, social/business, applications

- **Methodologies**
  - Database, information retrieval, signal and image processing, graphics, vision, human-computer interaction, machine learning, statistical modeling, data mining, pattern analysis, data fusion, social sciences, and domain knowledge for applications
Hierarchical Levels

High-level
Bridge the semantic gap, integration of context and content, hybrid (text and content) approaches

Mid-level
Active Learning, Incremental Learning

Low-level
Feature Extraction and Representation, Dimension Reduction and Selection.
Overview

Multimedia Information Retrieval

Content-based Video Retrieval

- Traditional Video Retrieval
  - Query-by-textual keyword

- Automatic Visual Concept Detection
  e.g., indoor/outdoor, Sky, Car, Building, US-flag
  Example concepts:
    Airplane, Building, Car, Crowd, Desert, Explosion, Outdoor, People, Vehicle, Violence

- Video Retrieval – Scene
  - How to recognize a scene? Context
    - Use Proto-Concepts to describe context
    - Machine learning to link context to concepts
Multimedia Information Retrieval

Content-based Audio Retrieval
- to search sounds by their features in the waveform, statistics, or transform domains
  - Speech, Music, Environment Audio, Silence

Applications
- Entertainment
  - Film making - searching sound effects
  - TV/radio studio - editing programs
  - Karaoke, music stores, or online shopping
    - query by humming the melody
- Audio/video archive management
  - Segmenting and indexing of raw recordings
  - Searching and browsing audio/video clips
- Surveillance
  - Monitoring criminal or emergent events
  - Film rating

- Short time energy
- Zero-crossing rate
- Pitch period
- MFCC
- Spectrogram
- LPC
- glass breaking, explosion, cry, shot, …
Top 10 Problems in MIR

- **Bridge the Semantic Gap**
  - high level concept (sites, objects, events) and low-level visual/audio features (color, texture, shape and structure, layout; motion; audio - pitch, energy, etc.).

- **How to Best Combine Human Intelligence and Machine Intelligence.**
  - Keep human in the loop, e.g. Relevance Feedback

- **New Query Paradigms**
  - Query by keywords, similarity, sketching an object, sketching a trajectory, painting a rough image, etc. Can we think of useful new paradigms?

- **Multimedia Data Mining**
  - Searching for interesting/unusual patterns and correlations in multimedia has many important applications, including Web Search Engines and dealing with intelligence data.
  - Work to date on Data Mining has been mainly in Text data.

- **How to Use Unlabeled Data**
  - Active learning, e.g., in Relevance Feedback
  - Label propagation, e.g., image/video annotation

Xiong, Zhou, Tian, Rui and Huang, “Semantic Retrieval of Video”, IEEE SP Mag., March 2006
Using Virtual Reality Visualization To Help
  - Can we use 3D audio/visual visualization techniques to help a user to navigate through the data space to browse and to retrieve?
  - e.g., 3D MARS

Incremental Learning
  - Change the parameters of the retrieval algorithms incrementally, not needing to start from scratch every time we have new data.

Structuring Very Large Databases
  - Researchers in audio/visual scene analysis and those in Databases and Information Retrieval should really collaborate CLOSELY to find good ways of structuring very large multimedia databases for efficient retrieval and search.

Performance Evaluation
  - Precision, recall, and more???

What Are the Killer Applications of Multimedia Retrieval?
  - e.g., medical multimedia, document management
Search in literature

Related Publications

**Journals**
- IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)
- IEEE Transactions on Circuits and Systems for Video Technology (CSVT)
- IEEE Multimedia (MM)
- International Journal on Computer Vision (IJCV)
- Pattern Recognition (PR)
- ACM Transactions on Knowledge Discovery from Data (TKDD)
- IEEE Transactions on Computational Biology and Bioinformatics

**Conferences**
- ACM Multimedia
- IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)
- International Conference on Pattern Recognition (ICPR)
- Others including ICME, ICIP, ICASSP, MIR, CIVR
Current Directions

- Web Image Search and Mining
- Image Annotation
- Affective Video Retrieval
- Information Fusion in MIR
- Integration of Context and Content for Multimedia Management
- Multimodal Emotion Recognition
Web Search

- Web Search 1.0 – Traditional Text Retrieval
- Web Search 2.0 – Page-level Relevance Ranking
- Web Search 3.0 – Object-level Structured Search

- Object Level Vertical Search (MSRA Libra: http://libra.msra.cn/)
- Live Product Search (http://products.live.com)
Image and video Annotation

- Photo and video sharing through the Internet has become a common practice.
  - flickr.com: 19.5 million photos (30% growth/month), 2005
  - Photo.net and airliners.net: millions of images

- Most image search engines relies on textual descriptions of the images, e.g., Google, Yahoo, MSN

  ![Google](image1)
  ![Yahoo](image2)
  ![MSN](image3)

- In general, people do not spend time labeling or annotating their personal photos or videos
Trends and Directions

Image and frame Annotation

Can computer do this?

- Building
- Sky
- Lake Landscape
- Tree

Image Annotation System

- A statistical model that can relate words to image features.
- Sketch an image, extract feature vectors.
- Descriptive words--top words ranked according to likelihood.

Recent work

Li, Wang (alipr.com 2006), Blei, Jordan (2003); Vasconcelos (UCSD-SML 2007), Zhang et al. (MSRA 2005), Li et al. (MSRA 2007)

Promising Direction:

Web Image Annotation – an integration of IR and Content Analysis
"Affective" Video Retrieval

Affective:
- “a feeling or emotion as distinguished from cognition, thought, or action”

Real Multimedia Retrieval
- Search for a subset of *nicest* holiday pictures to show them to friends
- Selecting *the most appropriate* background music for the given situation
- Search for *the most impressive* video clips
- Search for the *most appealing* photographs of one and the same content
- Search for all film comedies *I like most*

Alternative approach
Search for
- Mood
- Matches to users’ profile
  - Like/dislike, interest/no interest

Affective Video Retrieval
Trends and Directions

Information Fusion in MIR

**Fusion**: A merging of diverse, distinct, or separate elements into a unified whole (Merriam-Webster dictionary).

- **Feature Extraction Module**:  
  - Multiple features -> vectors  
  - Concatenated vector  
  - **Feature Fusion**: more discriminating hyperspace can be found in the new vector

- **Matching Module**:  
  - One type of classifiers for multiple features or  
  - Multiple types of classifiers for one feature or  
  - Both  
  - The output score can be combined

- **Decision Module**:  
  - The output decision of each classifier can be combined
Information Fusion in MIR

Two Forms

- **Multi-modality**
  - e.g. Video clip -> visual information, audio information, textural information
  - Multi-modality fusion occurs at feature extraction module.
  - Single source information may be represented by multiple features, e.g. Color image -> color, texture, shape

- **Multi-Classifiers (Ensemble of Classifiers)**
  - A set of classifiers are trained to solve the same problem
  - Applied on single or multiple source of information
  - A single type of base classifiers or
  - Different types of classifiers (Bayesian, K-NN, SVM)
Information Fusion in MIR

Fusion Schemes

- The prediction of multiple classifiers need to be integrated into one fused decision by *Fusion Scheme*.
  - The output of different classifiers need to be normalized.
- **Rule-based**
  - Decision is made by a simple operation on the output of all classifiers
  - e.g. Max, Min, Sum, Mean (Matching Module)
  - e.g. AND, OR (Decision Module)
- **Learning-based**
  - The output of all classifiers is fed into a *learning process* to obtain the final decision
  - e.g. Decision Tree, Neutral Network

No one is guaranteed to be the best empirically or theoretically.

Applications

- Multimodal Biometrics Fusion (face, fingerprint, iris, palmprint, voice, hand geometry)
- Audio/visual fusion for multimodal emotion recognition
Integration of Context and Content for Multimedia Management

from Carlson & Hatfield, 1992
Integration of Context and Content for Multimedia Management

- An increasing number of active research in this direction
- Crucial to human-human communication and human understanding of multimedia.
  - without context it is difficult for a human to recognize various objects,
- Enable that the (semi)automatic content analysis and indexing methods become more powerful in managing multimedia data
- Contextual information
  - e.g., cell ID for the mobile phone location, GPS integrated in a digital camera, camera parameters, time information, and identity of the producer
Integration of Context and Content for Multimedia Management

Topics of interest include:

- Contextual metadata extraction
- Models for temporal context, spatial context, imaging context (e.g., camera metadata), social context, and so on
- Web context for online multimedia annotation, browsing, sharing and reuse
- Context tagging systems, e.g., geotagging, voice annotation
- Context-aware inference algorithms
- Context-aware multi-modal fusion systems (text, document, image, video, metadata, etc.)
- Models for combining contextual and content information
- Context-aware interfaces and collaboration
- Novel methods to support and enhance social interaction, including innovative ideas like social, affective computing, and experience capture.
- Applications such as using context and similarity for face and location identification
- Context-aware mobile media technology and applications
- Using context to browse and navigate large media collections
Projects in High Impact

Photo & Video Search
- Home photo/video management
- Mobile photo/video search
- Face search

Human-centered Multimedia Search
- Search for U&C data
  (user preference, profile, and opinion)
- Social search
  (public relations, names, personalization)