Content-Based Multimedia Information Retrieval

Ishwar K. Sethi
Intelligent Information Engineering Laboratory
Department of Computer Science & Engineering
Oakland University
Rochester, MI 48309
Email: isethi@oakland.edu
URL: www.cse.secs.oakland.edu/isethi

Department of Computer Science & Engineering

Laboratory

Department of
Computer Science
& Engineering

Content?

Descriptive content

Subjective content

Behavioral reaction of a viewer to the image

Content-Based Information Retrieval (CBIR)

An inherently difficult problem because "what is actually in a document" is a function of both the document and the user. The ideal situation for perfect retrieval occurs when the document representation of the retrieval system and document representation of the user are in complete match.

Department of Computer Science & Engineering

Types of CBIR Queries

- Level 1
  - Find pictures with round red objects in the top left-hand corner
- Level 2 (Descriptive queries)
  - Find images containing multistory buildings
- Level 3
  - Find images showing tranquility

Department of
Computer Science
& Engineering

Current Content-Based Retrieval Methods

Department of
Computer Science
& Engineering

- Keyword-based retrieval (KBR)
- Similarity-based retrieval (SBR)

Keyword-Based Retrieval

- Good for finding images containing instances of desired objects (descriptive queries)
- Manual cataloging
- High expressive power
  - *Can be used to describe any aspect of image content at various levels of complexity*
- Subject to user differences
  - *Two people choose the same main keyword for a single well-known object only about once in five times*

Department of
Computer Science
& Engineering

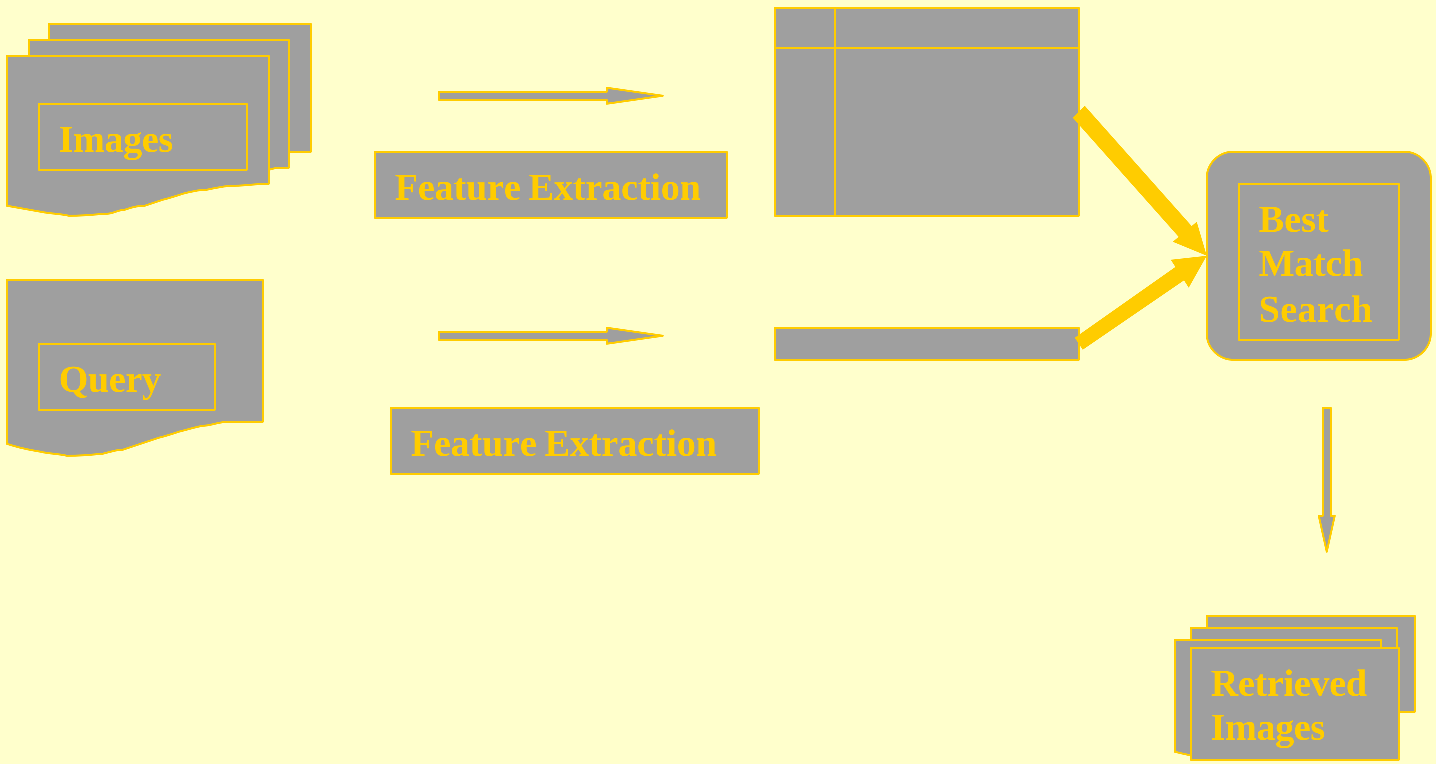
Department of
Computer Science
& Engineering

Similarity-Based Retrieval

Avoids issues related to manual cataloging
Suitable for computerized indexing
Able to capture the compositional aspects to a limited extent
Good for Level 1 queries

Department of
Computer Science
& Engineering

Similarity-Based Retrieval



An Example of SBR

Department of
Computer Science
& Engineering

Major Limitation of the SBR Approach

Signal versus descriptive/semantic content similarity (Semantic gap)

Department of
Computer Science
& Engineering

How to Reduce the Semantic Gap?

- Stuff detectors
- Image category detectors / feature associations
- Exploiting other information sources
  - Surrounding text / image captions
  - Associated audio
  - Cross-modal association

Department of
Computer Science
& Engineering

Stuff Detectors

Stuff detectors are object detectors. Current computer vision methods allow to build a small set of special detectors, each designed to detect the presence of a particular type of “stuff.” Examples of some stuff detectors include

- faces
- traffic signs
- trees

Department of
Computer Science
& Engineering

Face Detector

Department of
Computer Science
& Engineering

Traffic Sign Detector

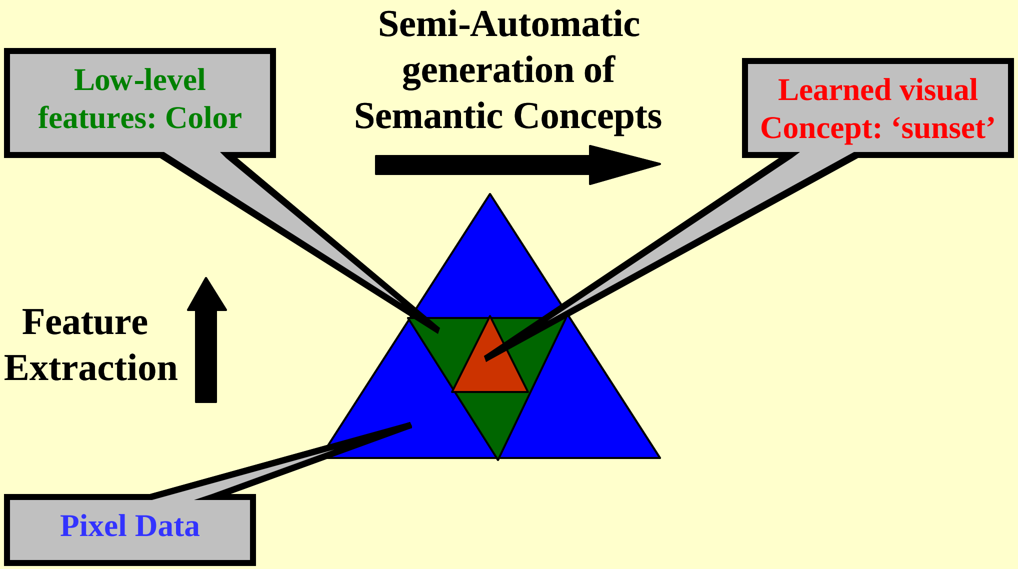
Image Category Detectors

These detectors try to determine the broad category of image content by building image classifiers. These detectors are different from stuff detectors which locate specific types of objects within an image. Here, the image as a whole is assigned a descriptive keyword.

Department of
Computer Science
& Engineering

Department of
Computer Science
& Engineering

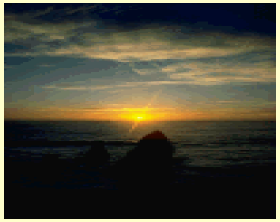
Image Category Learning

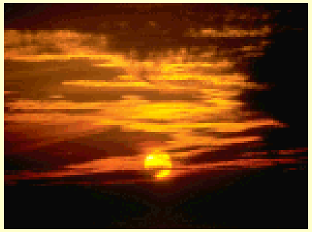


The relationship between image data, low-level features, and high-level concepts (image categories) can be visualized using the triangle relationship between data, information, and knowledge: low-level features (information) are extracted purely from pixel data, and knowledge (learned visual concepts) is discovered from the most important low-level features and image contexts.

Department of
Computer Science
& Engineering

An Example of Image Category Classification






Sample images classified as '*sunset*' by a rule-based image classifier, eID system

Codebook Based Image Category Detection

- Good for *mass noun entities*, for example grass, water, sand etc.
- Entity specific codebook designed through vector quantization
- A confidence value is attached to each codeword in the entity specific codebook
- Image category is decided by encoding a given image through different entity specific codebooks and integrating the resulting confidence values

Department of Computer Science & Engineering

Department of
Computer Science
& Engineering

Vector Quantization Based Image Category Classifier

Smoke Agent

Fire Agent

Grass Agent

Sky Agent

Water Agent

Oakland University

Department of Computer Science & Engineering

Exploiting Text Surrounding Images

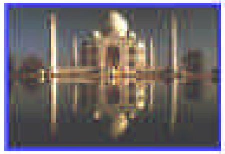
- Keywords extracted from text surrounding images can provide a way of reducing semantic gap
- The image search engine Google, for example, has cataloged over 450 million images using the surrounding text to extract keywords

IIE Laboratory

Department of
Computer Science
& Engineering

Google Example for "Taj Mahal"

tajmahal.jpeg
738 x 493 pixels - 78k
www.suite101.com/.../4205/
files/tajmahal.jpeg



tajmahal.jpg
384 x 256 pixels - 20k
www.theindiatravel.com/.../
photo/tajmahal.jpg

tajmahal.jpg
1765 x 1177 pixels - 85k

tajmahal.jpg
800 x 600 pixels - 75k

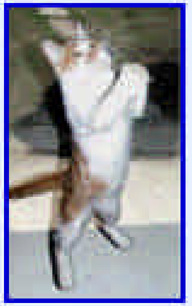
Keyword = Taj Mahal, Source = Google Image Finder

Google Result for “Prayer”

wi hands in **prayer** with c...
640 x 480 pixels - 354k
www.worshipimages.com/images/

prayer.gif
576 x 841 pixels - 112k
shc.stanford.edu/.../

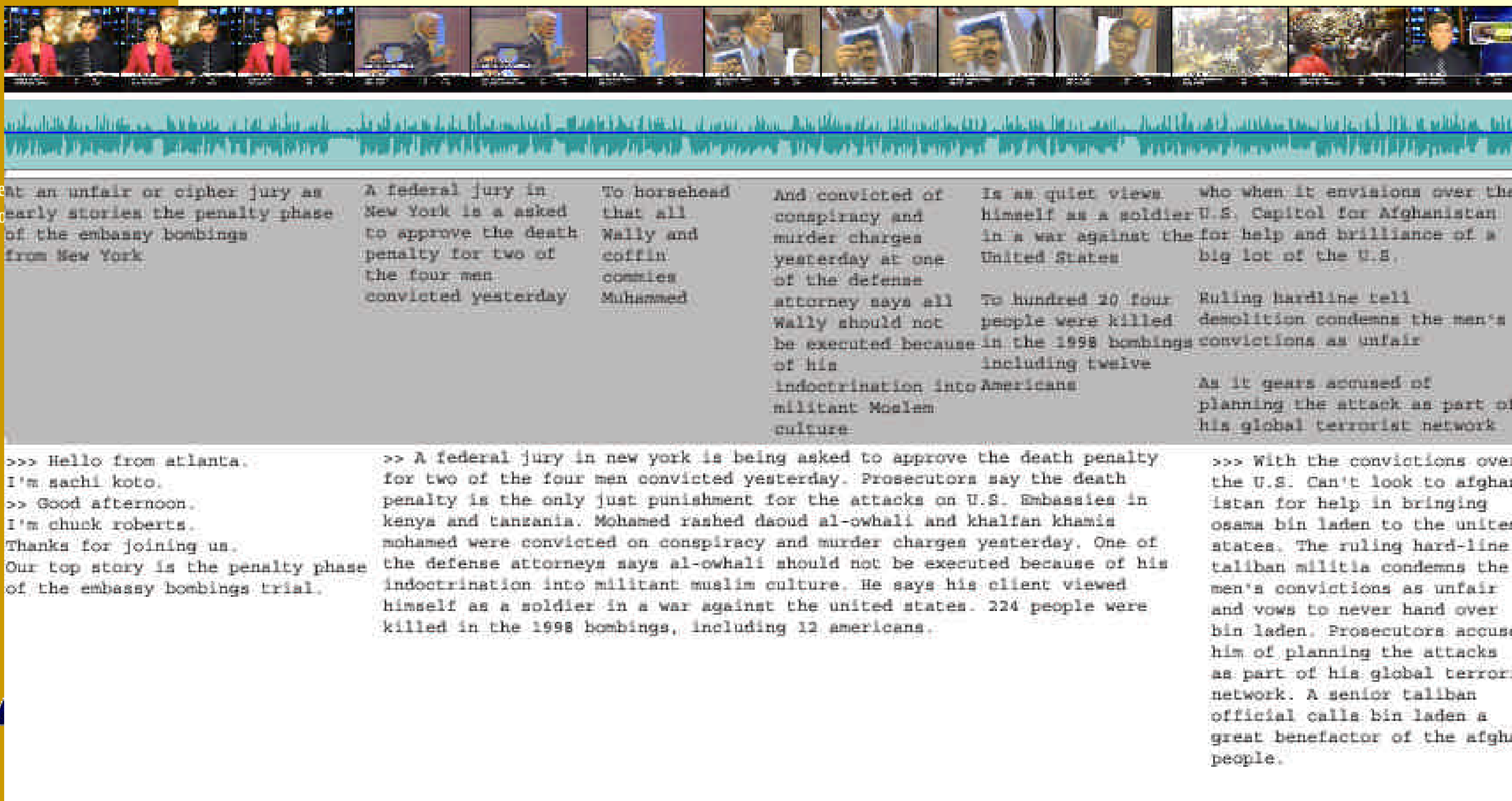
prayer.jpg
250 x 200 pixels - 6k
www.womanlinks.com/cards/prayer.jpg



prayer.jpg
254 x 443 pixels - 27k
skole.hinet.hr/ss-rovinj-503/skola/prayer.jpg

Department of Computer Science & Engineering

Information Sources in a Multimedia Stream



Video Analysis for CBIR

- What should be the analysis level?
  - A frame? A shot? A scene?
- Scene components
  - Objects (who), action or event (what), and place or context (where)
- Compositional components
  - Camera shot, angle, and movement
- Subjective components
  - Emotion and mood

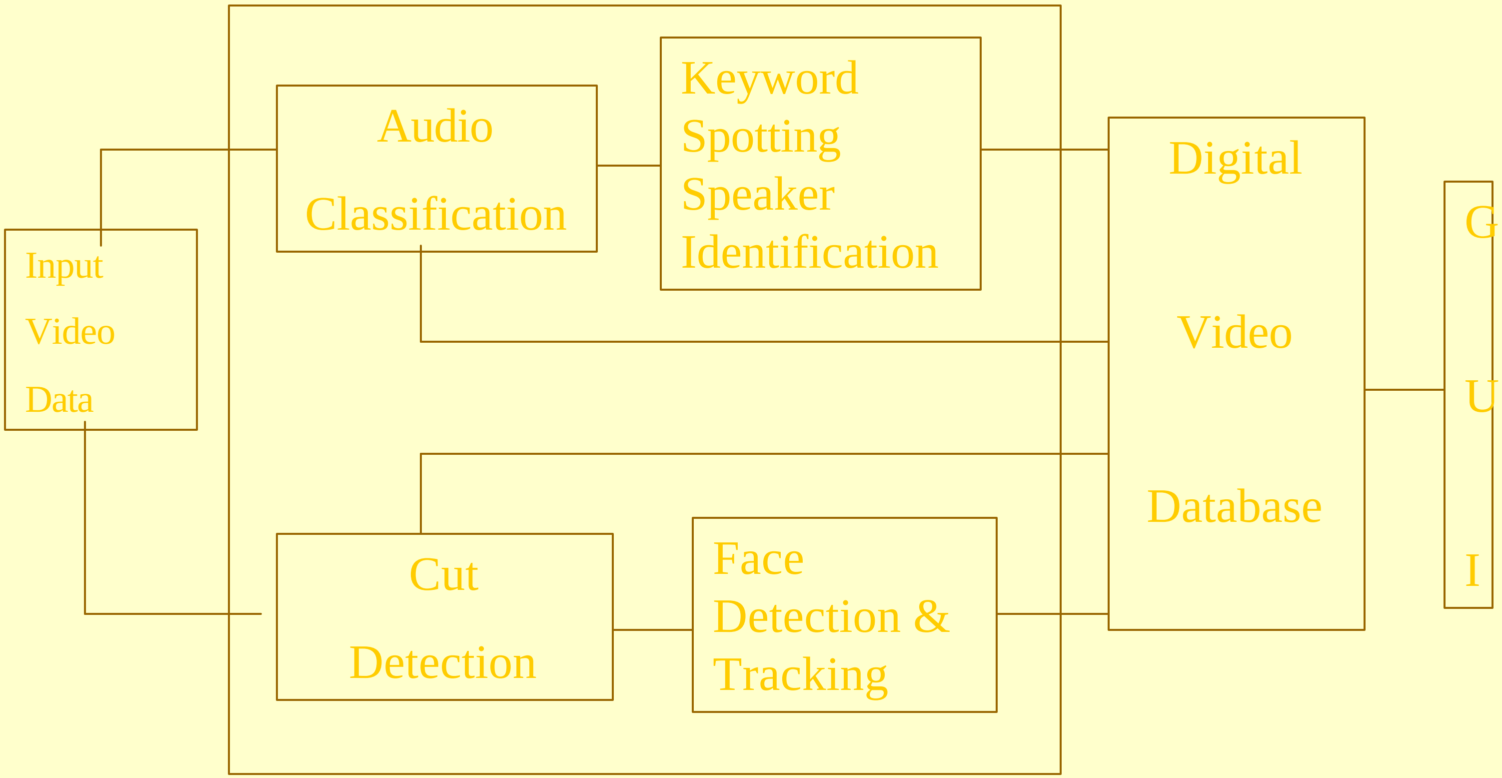
Department of Computer Science & Engineering

Department of
Computer Science
& Engineering

Integrated Analysis Approach for Video

- Video and image analysis
  - face detection, tracking, and recognition
- Audio analysis
  - audio segmentation and classification
  - speech/speaker recognition
  - text understanding
- Closed caption text analysis
- Transcript understanding

Partial Block Diagram of the Integrated System



Department of Computer Science & Engineering

Audio Analysis for Video Indexing

- Audio segmentation and classification
- Speaker identification
- Keyword spotting
- Speech recognition
- Text understanding

Department of
Computer Science
& Engineering

Cross-Modal Retrieval

Locate or retrieve documents of all modalities in response to a query in any modality

Department of Computer Science & Engineering

Opportunistic Vs Cross-Modal Integration

- Opportunistic Approach
  – The data from different modalities is processed independently and the results are used/merged on a *need* basis
- Cross-Modal Association Approach
  – The data from different modalities is processed together to discover and exploit associations between different modalities

Department of Computer Science & Engineering

Cross-Modal Association Approach

Department of
Computer Science
& Engineering

- Operates in the joint feature space
- Works by identifying and measuring intrinsic associations between different modalities
  - For example, facial features with speech
  - Uses feature sets that preserve/represent best such relationships

Department of
Computer Science
& Engineering

Work Related to CMA

- FaceSync by Slaney and Covell (NIPS 2000)
  - Synchronizing visual and speech streams using canonical correlation
- Monologue detection by Iyenger and Nock (ICASSP 2003)

Oakland University
Department of Computer Science & Engineering
IIE Laboratory

Possible CMA Approaches

- Model-based approaches
  - Gaussian distribution, linear correlation models, etc.
  - Learn fast and provide best results when using appropriate models
- Model-free approaches
  - Neural networks
  - Require little prior knowledge

CMA Using Linear Correlation Models

- Linear correlation model
  - Appropriate model for many applications when analysis time window is relatively short
- Possible models
  - Latent semantic indexing (LSI)
  - Cross-modal factor analysis (CFA)
  - Canonical correlation analysis (CCA)

Department of Computer Science & Engineering

Latent Semantic Indexing

Department of
Computer Science
& Engineering

- Popular in text information retrieval as an effective tool to relate keywords
- Extended to the multimedia domain, for example, to discover semantic associations between low-level multimedia features and keywords/captions
- Provides dimensionality reduction
- LSI may not provide the best representation of cross-modal relationships as the computation of the linear transformation is affected by intra-class distribution

Department of Computer Science & Engineering

CCA: A Possible Solution for CMA

- The nature of CMA is to examine the relationships between two feature subsets
  - distribution of patterns and noise within each subset should not be a factor
- With linear correlation model, the problem is to find the optimal transformation space
  - best represents the coupled patterns between two subsets
- CCA optimization criteria
  - Given coupled samples from two feature subset: X and Y, we seek A and B that

$$\max\{correlation(XA, YB) = correlation(\tilde{X}, \tilde{Y})\}$$

Oakland UNIVERSITY

Department of
Computer Science
& Engineering

IIE Laboratory

Canonical Correlation Analysis

$$A = C_{xx}^{-1/2} U \qquad B = C_{yy}^{-1/2} V$$

Where,

$$C_{xx} = E\{(X - m_x)(X - m_x)^T\}$$

$$C_{yy} = E\{(Y - m_y)(Y - m_y)^T\}$$

$$C_{xy} = E\{(X - m_x)(Y - m_y)^T\}$$

$$K = C_{xx}^{-1/2} \cdot C_{xy} \cdot C_{yy}^{-1/2} = U \cdot S \cdot V^T$$

Restriction:
no two features in each subset are correlated

Oakland University
Department of Computer Science & Engineering
IIE Laboratory

Cross-Modal Factor Analysis: Another Possibility

- Optimization criteria
  - we seek transformation A and B that minimize

$$\left\| XA - YB \right\|_F^2$$

We can prove that this is equivalent to maximizing :

$$trace(XAB^T Y^T)$$

$$\begin{cases} A = S_{xy} \\ B = D_{xy} \end{cases} \quad \text{where } X^T Y = S_{xy} \cdot V_{xy} \cdot D_{xy}$$

Cross-Modal Factor Analysis

- Transform X and Y using A and B

$$\begin{cases} \tilde{X} = X \cdot \tilde{A} \\ \tilde{Y} = Y \cdot \tilde{B} \end{cases}$$

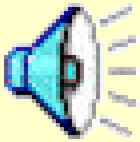
- Pearson correlation or mutual information can then be used

Department of Computer Science & Engineering

Cross-Modal Factor Analysis

X samples

Y samples



CFA Training

A

B

First 7 most important vectors of A reshaped to corresponding visual location:

A2 A3 A4 A5 A6 A7

Department of
Computer Science
& Engineering

Department of
Computer Science
& Engineering

CFA vs. CCA

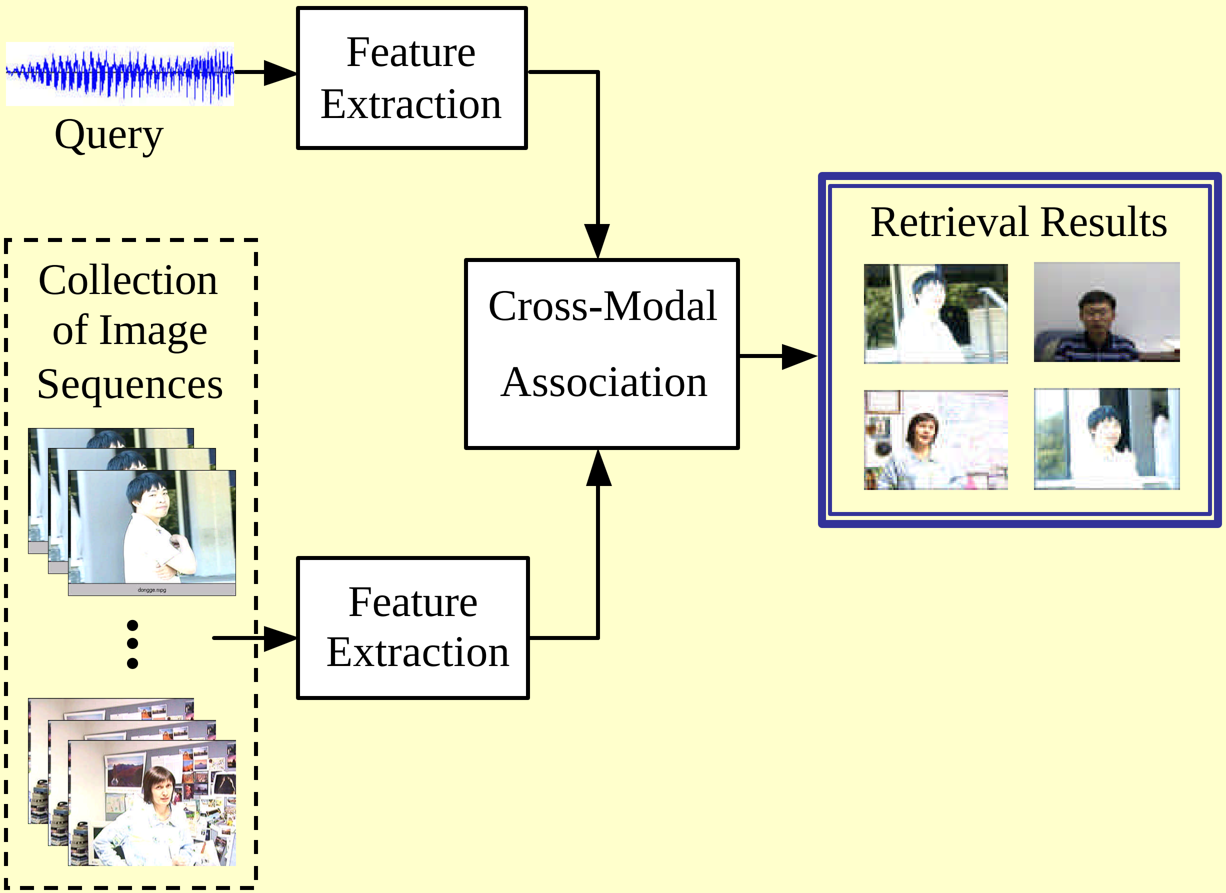
- Transformation matrixes given by CFA are orthogonal, while not necessary for CCA
- CFA is in favor of correlation patterns with high variations, while CCA is more sensitive to patterns with low variations due to the calculations of $C_{xx}^{-1/2}$ and $C_{yy}^{-1/2}$
- CFA does not have the de-correlation restriction on the features

Advantages of Cross-Modal Retrieval

Department of Computer Science & Engineering

- Greater choice for input modalities
  - generating and sending query of a more appropriate modality
- Handle absent (corrupted) modalities
- More flexible browsing of multimedia databases
- Potential to combine with existing single-modality approaches

System Structure of CMR



Department of Computer Science & Engineering

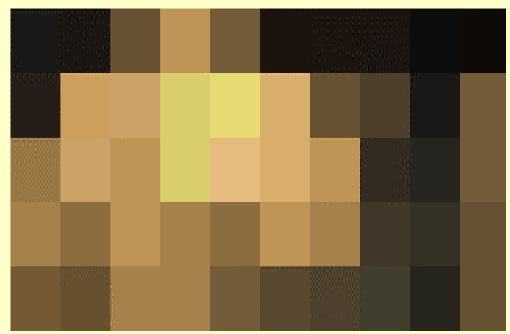
Example 1: Retrieval of Explosion Scenes

Department of
Computer Science
& Engineering

- Audio query - 4 second explosion clips
- Visual database: 452 explosion clips and 3870 non-explosion clips
  - many are low quality without soundtracks
- Audio features
  - 12 MFCCs

Example 1: Retrieval of Explosion Scenes (2)

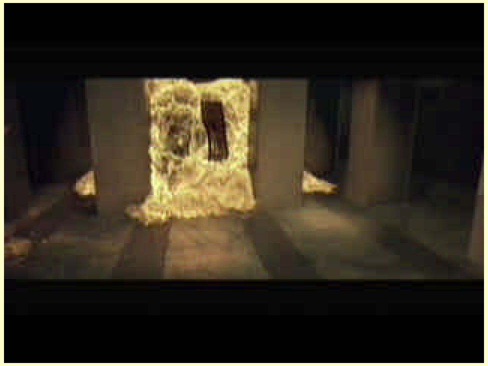
- Visual features:
  - 150 HSI area-peak values from 5x10 overlapped image blocks



- Only 8 most important features after the transformation are kept

Department of Computer Science & Engineering

Retrieval examples of explosion scenes



Department of
Computer Science
& Engineering

Oakland University

Department of Computer Science & Engineering

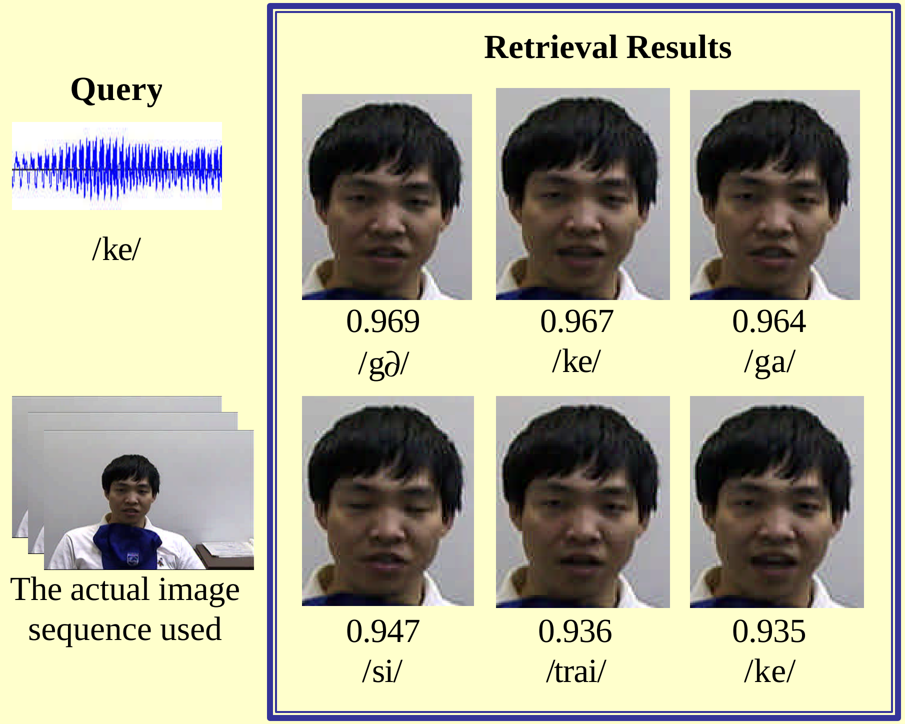
Performance Comparison

Hit Rate	**CFA**	**CCA**	**LSI**
Top 5	62%	61%	21%
Top 10	41%	42%	21%
Top 20	37%	32%	20%

IIE Laboratory

Example 2: Retrieval of Talking Faces

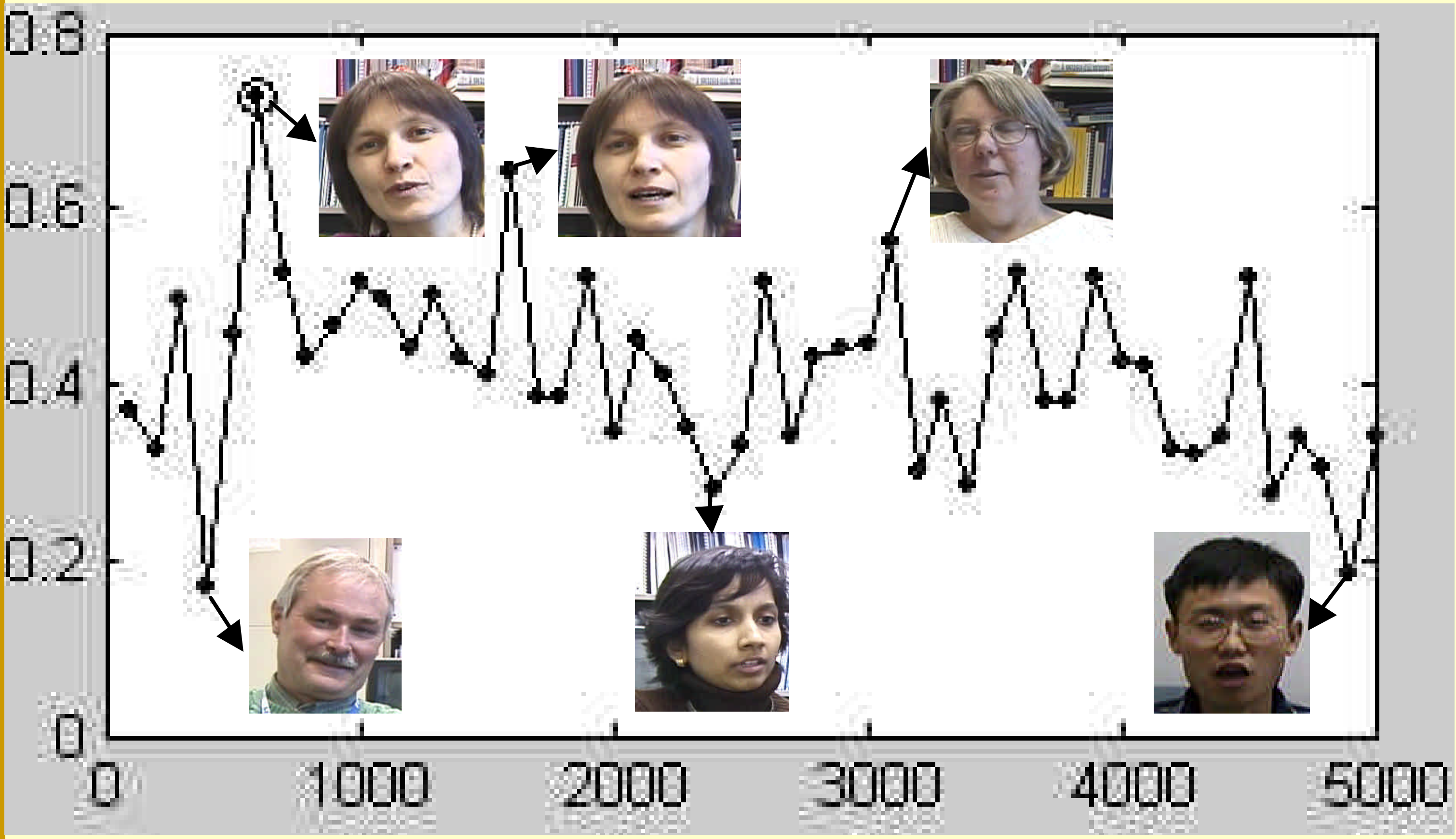
- Audio query - single syllable audio clip
  - 12 MFCCs as audio features
- Visual features are 40x32 pixels from detected face areas



Department of
Computer Science
& Engineering

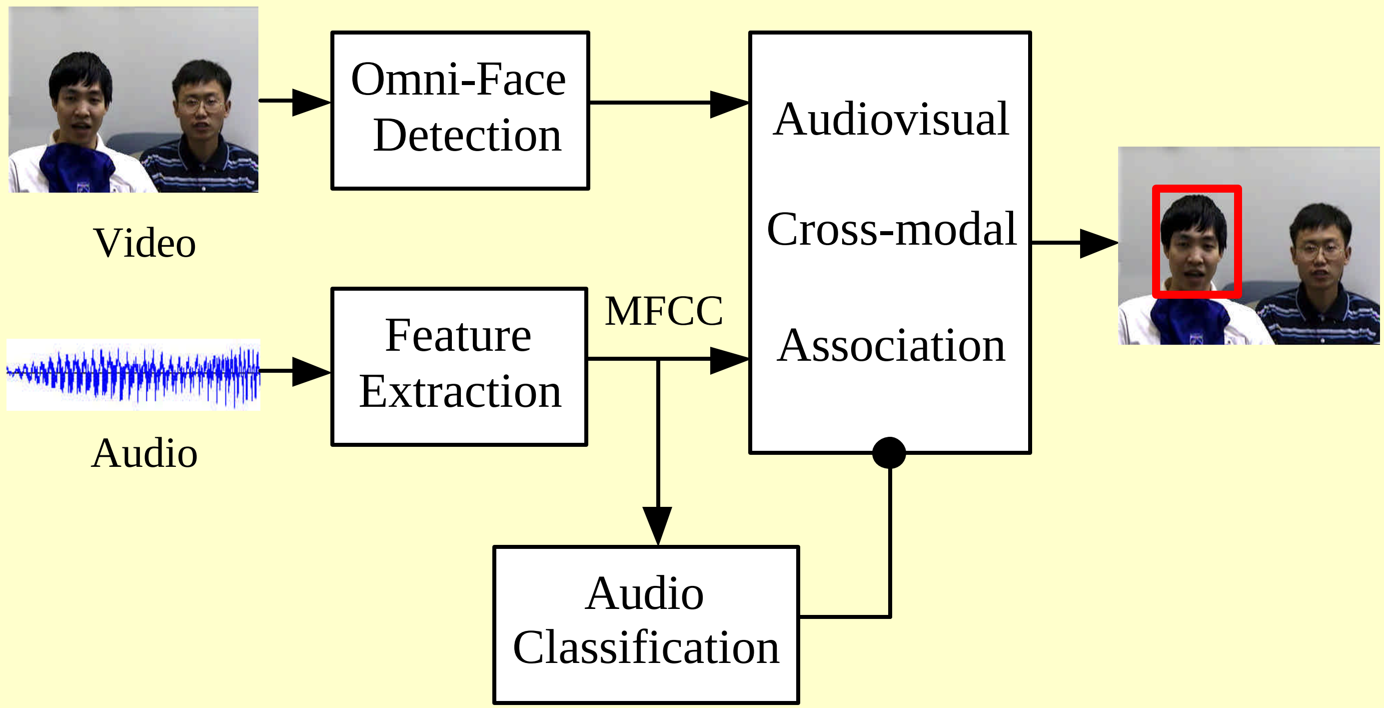
Department of
Computer Science
& Engineering

Example 2: Retrieval of Talking Faces



Department of
Computer Science
& Engineering

Talking Head Detection



Visual features: 40x32 image pixels
Audio features: 12 MFCCs

Performance Comparison

Department of
Computer Science
& Engineering

- Detection accuracy:
  - CFA: 91.1%
  - CCA: 73.9%
  - LSI: 66.1%
- CCA is more prone to noise due to its sensitivity to patterns with low variations

Oakland University

Department of Computer Science & Engineering

IIE Laboratory

Summary & Conclusion

- Level 1 queries are no problem
- Level 2 queries can be dealt with somewhat success using multiple information sources, image classifiers. Emerging techniques such as *active learning* are likely to play a greater role
- CMA offers a systematic approach for exploiting associations and extending the capabilities of multimedia information retrieval systems

Department of
Computer Science
& Engineering

Acknowledgement

- Gang Wei (Accenture)
- Dongge Li (Motorola)
- Daniela Stan-Raicu (DePaul)
- Victor Kulesh (Kulesh Software)
- Mingkun Li (Oakland University)