Multimedia Search

Lynn Wilcox

What is Multimedia?

Definition from the ACM Special Interest Group on Multimedia Retreat in 2003

- More than one media (text, images, audio, video) that are correlated
- Examples:

  Time correlated: Video with text transcript of the audio

  Spatially correlated: Images on a page with associated text

A less strict definition:

Not "Just" Text

- Images
- Audio
- Video

Multimedia Search Outline

Text Search

- Keywords

Image Search

- Search based on tags (FlickR, FaceBook)
- Search based on surrounding text (Google)
- Content based search
  - *Using image features*
  - *Using faces*

Audio Search

- Search based on metadata (iTunes)
- Content based search (MuscleFish, Foote)

Video Search

- Search based on text (Google/UTube)
- Search based on associated media (Lectures with slides)
- Search based on content (TrecVid News Search)

MediaMagic

Text Search

Documents (Web pages) represented by words

Inverted index links keywords to the documents that contain them

Keyword query retrieves documents containing that word

Inverted Index Example

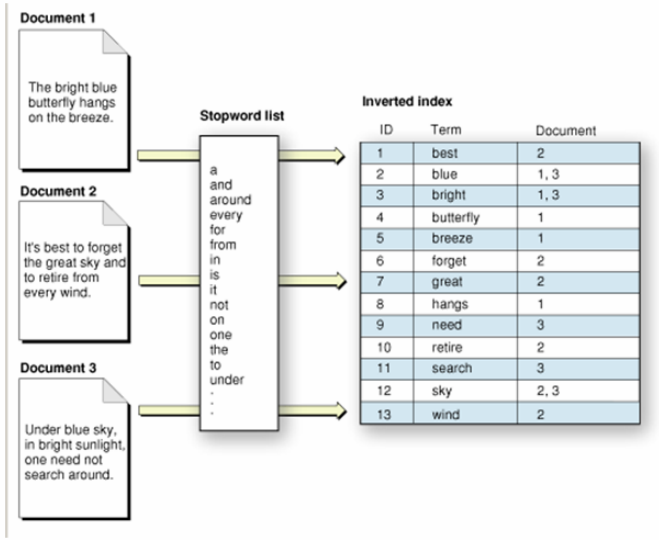


Image from http://developer.apple.com /documentation/UserExperience/Conceptual/SearchKitConcepts/searchKit_basics/chapter_2_section_2.html

Image Search - Tags

Search over tags associated with images

- Users manually add Tags to images
  - *FlickR*
  - *FaceBook*
- Find images with tags that match the query keyword

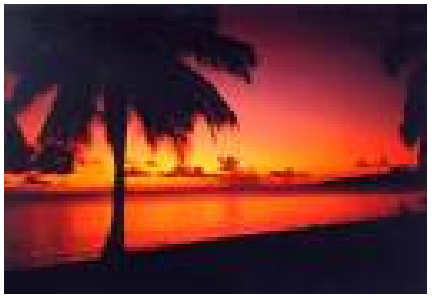
Limitations

- Tags require human effort to create
- Tags may be wrong

Image Search - Text

Use text associated with images for search

- Search web for images
- Use surrounding text
  - *Text in URL for image filename*
  - *Text in HTML on page*
- Same as text search



Sunset at Rocky Point

Example: Google Image Search for "Sunset" gives

- Sunset at Rocky Point in Australia
- Sunset Beach, Oahu
- Frank Smiles at Sunset

Frank Smiles at Sunset

Sunset Beach

Because the keyword "Sunset" was in the title of all these images

Image Similarity Search

Query is an image

Search finds similar images

Similarity is defined by features of the image

- Color Content
  - *Color Histogram*
  - *Color Corellogram*
- Image descriptors
  - *Gradients at image keypoints*
  - *Quantize for "Visual words"*
- Faces
  - *Detection*
  - *Recognition*

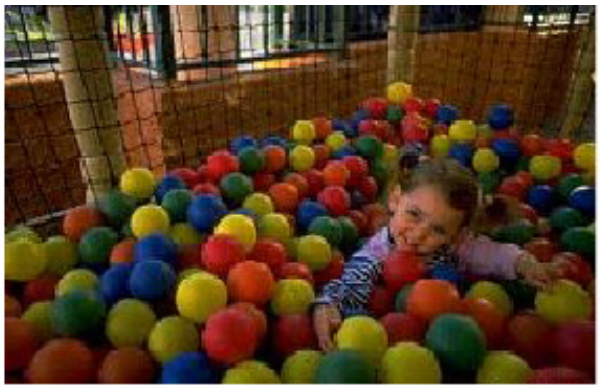
Query Image

Search Results

Image Search – Similar Color Content

Color Histogram

- Distribution of pixel colors in image
- No spatial information
- Similarity based on histogram distance



Color Correlogram

- Color histogram as a function of distance between pixels

  Multiple color histograms – one for each distance

- Distribution of pixel color plus spatial information
- Similarity based on correlogram difference

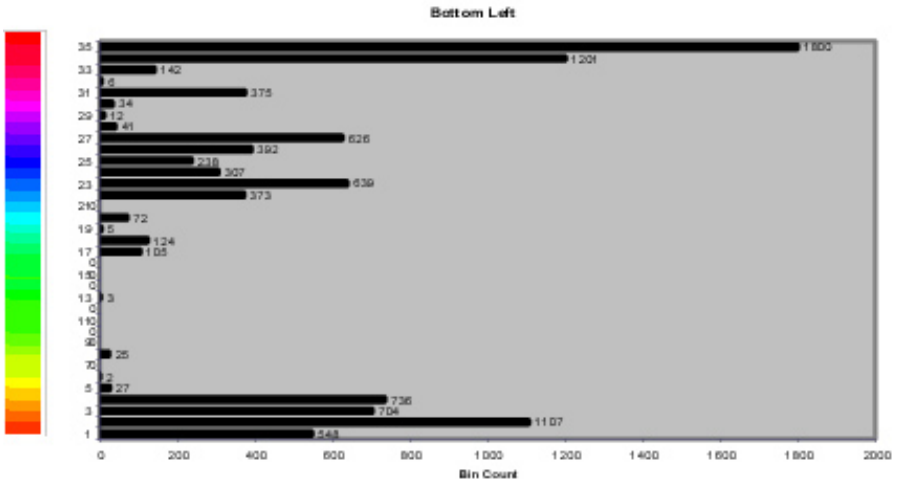
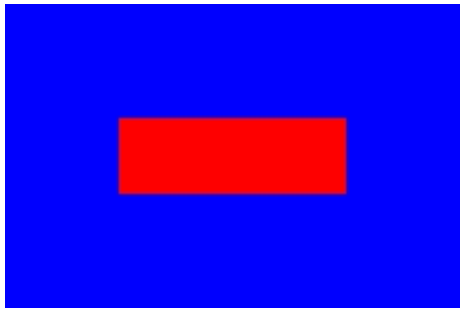


Image Search Results – Color Content

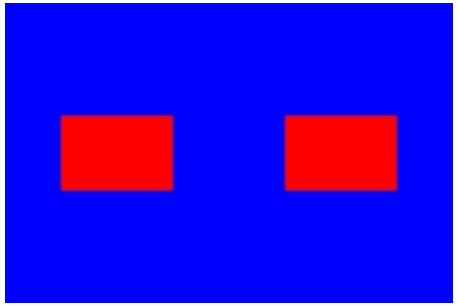
Comparison of Image Retrieval

- Color Histogram
- Color Correlogram

Correlograms are better for image retrieval



Images with identical color histograms but different correlograms



Images with different color histograms but similar correlograms

Image Search – Image Descriptors

SIFT Features

- (Scale Invariant Feature Transform) Features
- (2004: David Lowe, UBC)

Select keypoints regions in image from extrema in scale space

- Different images have different numbers of keypoints

Compute feature vectors X for each keypoint region

- Feature vectors from histogram of gradient directions near the keypoint
- SIFT features X are 128-dimensional vectors

Image described by N SIFT features

- Features are X_1,….X_N
- N is different for different images

Image Search – Visual Words

Quantize SIFT features to create "visual words" to represent images

- (2006: Lienhart, University of Augsburg & Slaney, Yahoo!)

Cluster SIFT features of representative images

- Features X are in 128-dimensional space
- Generate W clusters
- Clusters define "visual words"
- All features in same cluster are the same "visual word"

To compute visual words describing an image

- Compute N SIFT $X_1,\ldots X_N$ features for the image
- Find nearest cluster center (codeword) to each features X_j
- These clusters define the visual words for the image

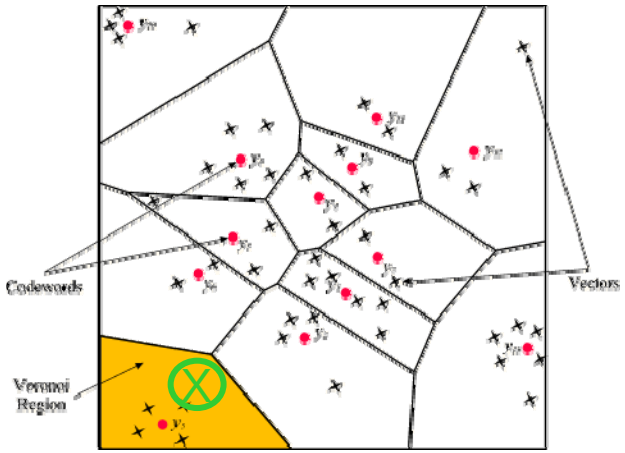


Image Feature ⊗

Visual Word 1

Image Retrieval – Visual Words

Image is described by it's visual words

- Just like a document is described by the text words

Create image index

- Compute visual words for all images
- Create a visual word index into the images

Compute visual words for query image

- Use query words for retrieval

Just like text!

- Except the visual words aren't quite as meaningful

Image Search –Faces

Face Detection

- Find faces in images
- Search for all images with faces
- Ex: Google advance search for images with faces
- Good results!

Example:

- FXPAL Photo Application (2004: Girgensohn et al.)

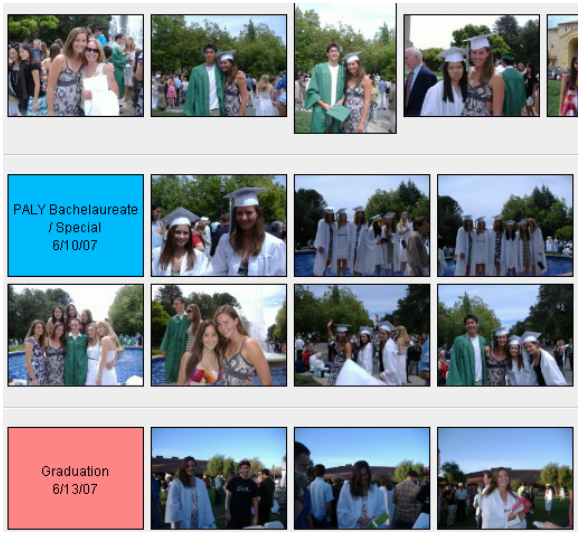


Photo Collection

Faces in Photo Collection

Face Detection

Image Search –Faces

Face Recognition

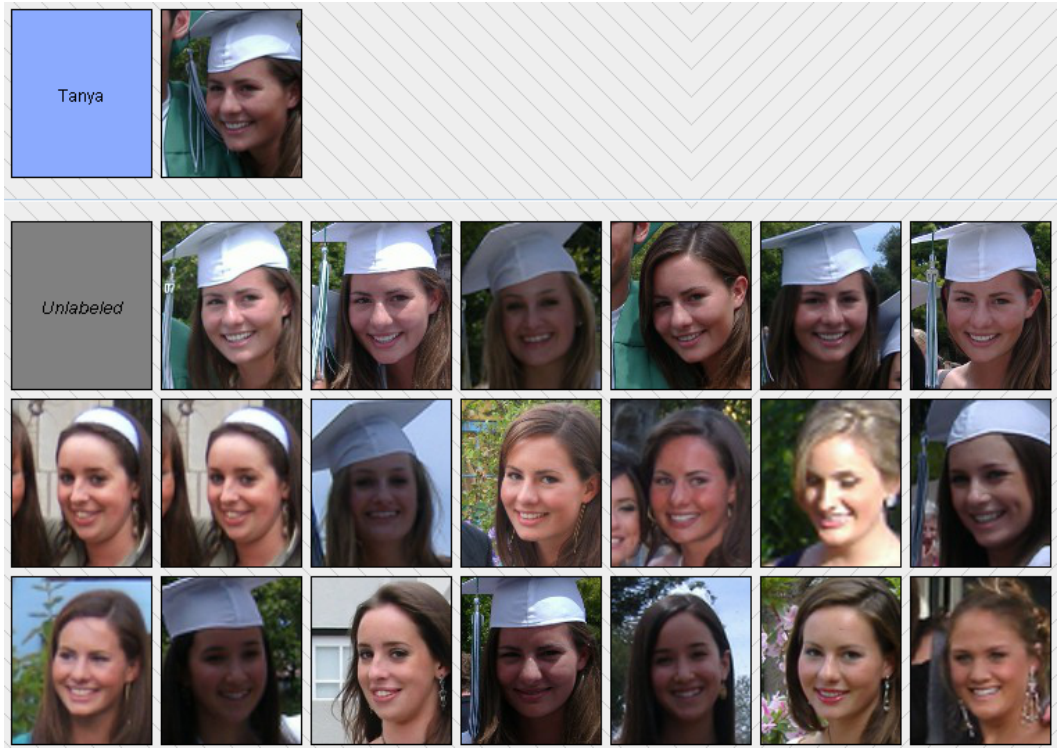
- Search for all images of a particular person
- Bad results!

Face Similarity

- Similarity search based on face features
- Use face similarity to help manually label faces
- Good results!

User Interface for Labeling Faces

- Drag face to label



Audio (Music) Search – Text

Search text fields

- Title
- Artist
- Album
- Genre

Example

- iTunes

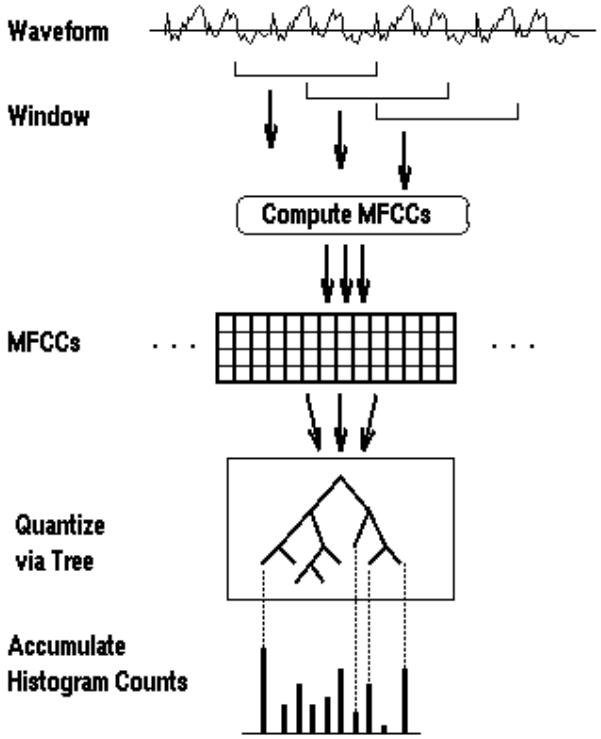
Audio (Music) Search - Sound

Find similar sounding music

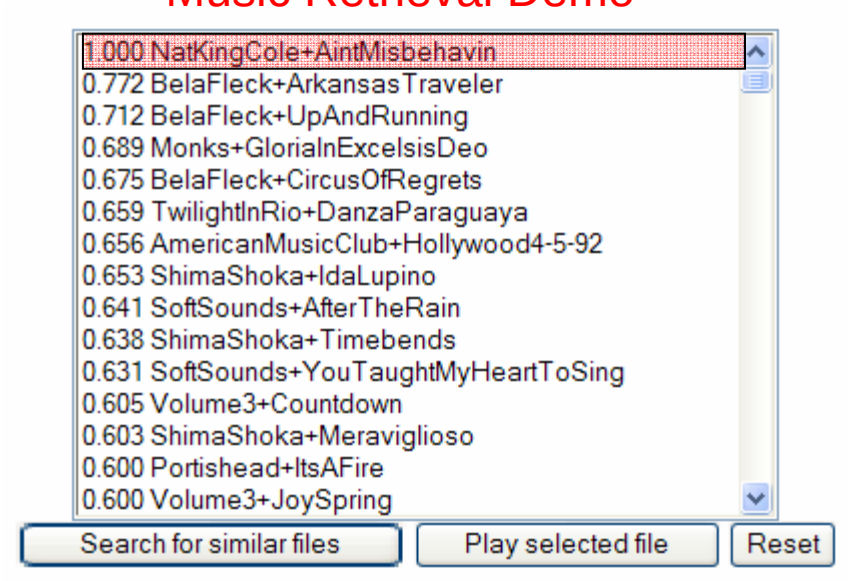
- Compute spectral feature vectors (MFCC)
- Quantize features to create audio histogram
  - *Audio histogram describes sounds*
  - *Order of sounds is lost*

Example

- 1997: Jon Foote, FXPAL
- Similarity of Nat King Cole and Gregorian Chant



Music Retrieval Demo



http://www.rotorbrain.com/foote/musicr/

Video Search – Whole Video

Search for an entire video

- Search using surrounding text

Example: Google/YouTube

- Search for sunset

Sunset - Nitin Sawhney
4 min - Nov 17, 2006 - ★★★★★ (39 ratings)
Nitin Sawhney**...Sunset** Very best video of Nitin Sawhney**...Sunset**
http://www.youtube.com/watch?v=nj6JiXjErTI
Watch video here - Related videos

Sunset at Cafe Del Mar 2004
6 min - Oct 17, 2006 - ★★★★★ (80 ratings)
Mar 2004...A realy Cool **Sunset** at Cafe Del Mar Ibiza.I overlaid two songs on this video, the first is Water in Motion by sonic
http://www.youtube.com/watch?v=vs1yzMNxhGk
Watch video here - Related videos

The Kinks - Waterloo **Sunset**
3 min - Jul 21, 2006 - ★★★★★ (429 ratings)
Waterloo **Sunset**...video of the kinks performing live version of waterloo **sunset**...The Kinks Waterloo **Sunset** water loo music
http://www.youtube.com/watch?v=fvDoDaCYrEY
Watch video here - Related videos

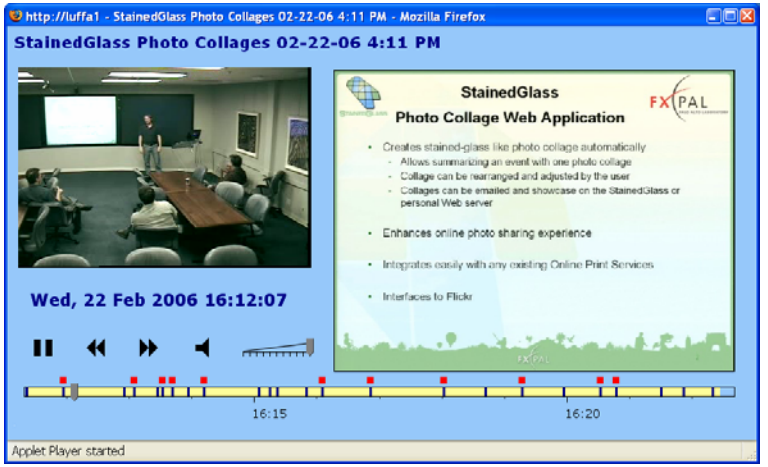
Video Search – Lecture Video

Search for a segment of a lecture

- Collection of lecture videos (maybe for a class)
- Find just that part of a lecture that you want to watch

Indexing Method

- Capture lecture audio and video
- Capture presentation material
- Extract text from presentation material
- Capture time correlations
- Segment the video based on slide change
- Create keyword index of segments of the video associated with each slide



Search Method

- Keyword search
- Play video starting at the relevant segment

Video and Audio Capture

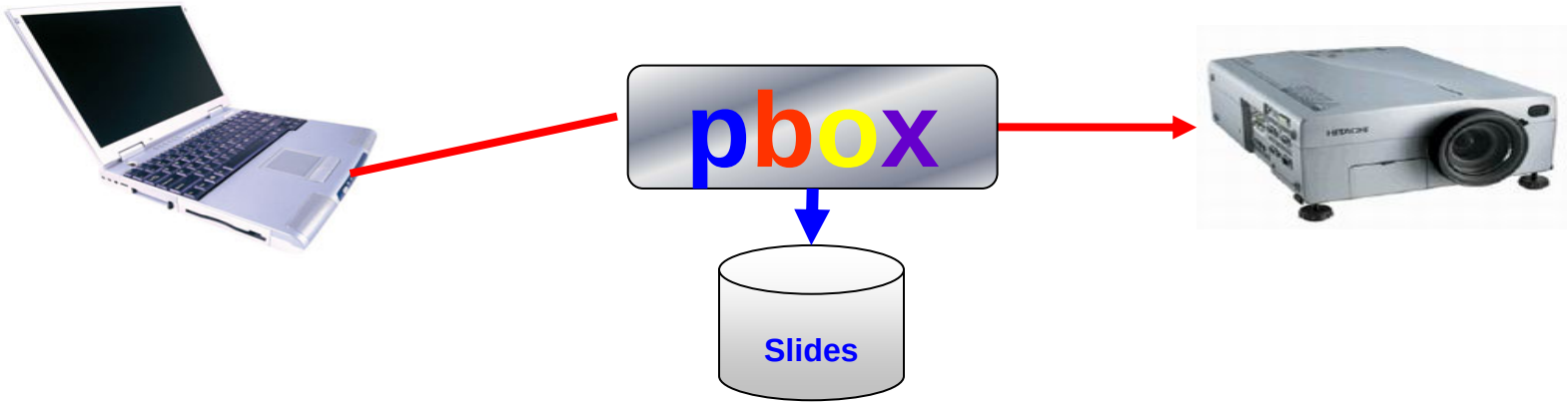
Automatic camera control

- Multiple cameras capture meetings from different angles
- Select camera with best view
- Pan and zoom to region of interest

Region of interest determined using

- Audio source
- Motion tracking
- Learning based on human operator

Presentation Capture



Capture Slide Images

- ProjectorBox (PBox): Denoue and Hilbert FXPAL
- Insert PBox in RGB stream between PC and projector
- Capture slides images and time stamps
  - *Capture slide images at a fixed rate*
  - *Only keep distinct slide*

Capture Text from Slide Images

- OCR slide images from PBox to get words
  - *Optical Character Recognition (OCR) to convert text image to electronic text*

Synchronize clocks of presentation and video capture devices

Video Search – Segments of Lecture Video

Video and Audio

- Sequence of frames (images) with audio
- 30 frames/second

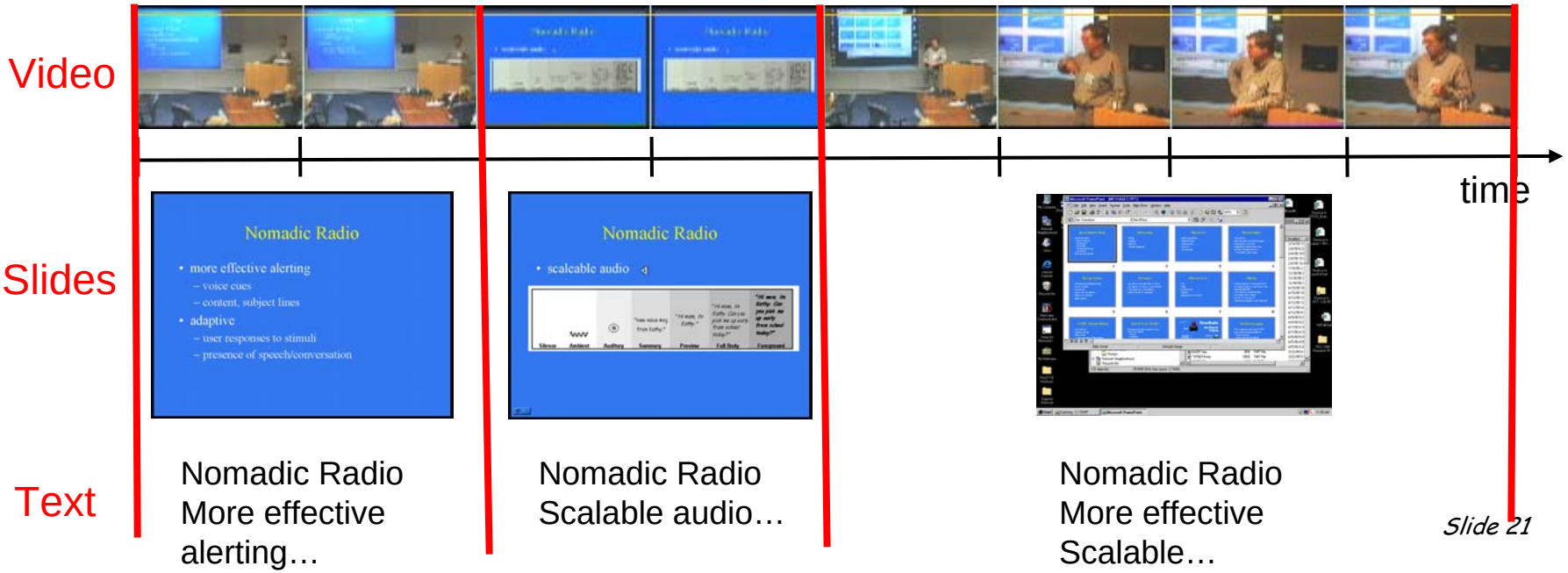
Presentation Slides

- Slide images
- Time-correlated with video

Text from Presentation Slides

Segment Video

- New segment when slide changes
- Video associated with a slide



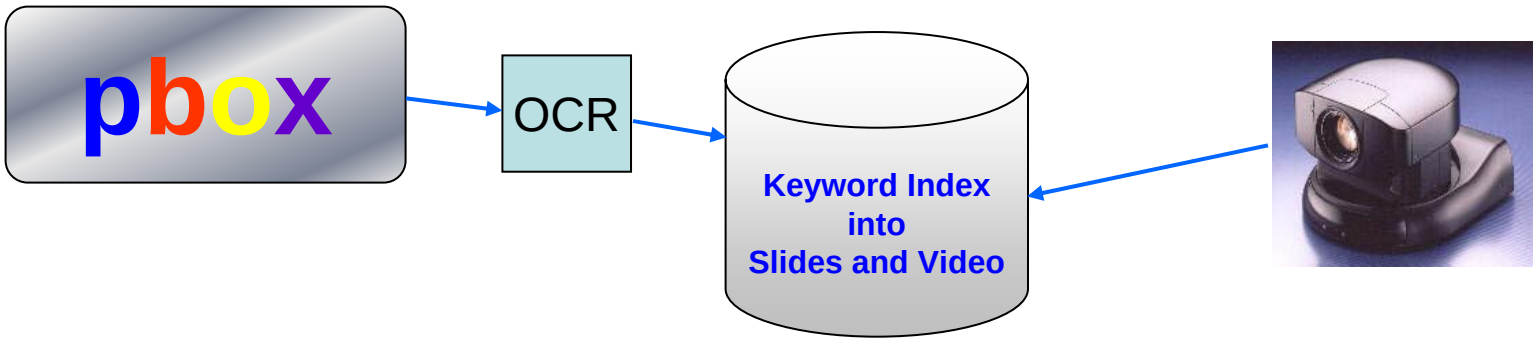
Lecture Video Index

Create index into video segments associated with each slide

- Index each slide in video based on text

Search

- Keyword search locates relevant slide
- Play video at starting time for that segment



Example: Find Presentation

Find Presentation on Photo Organization

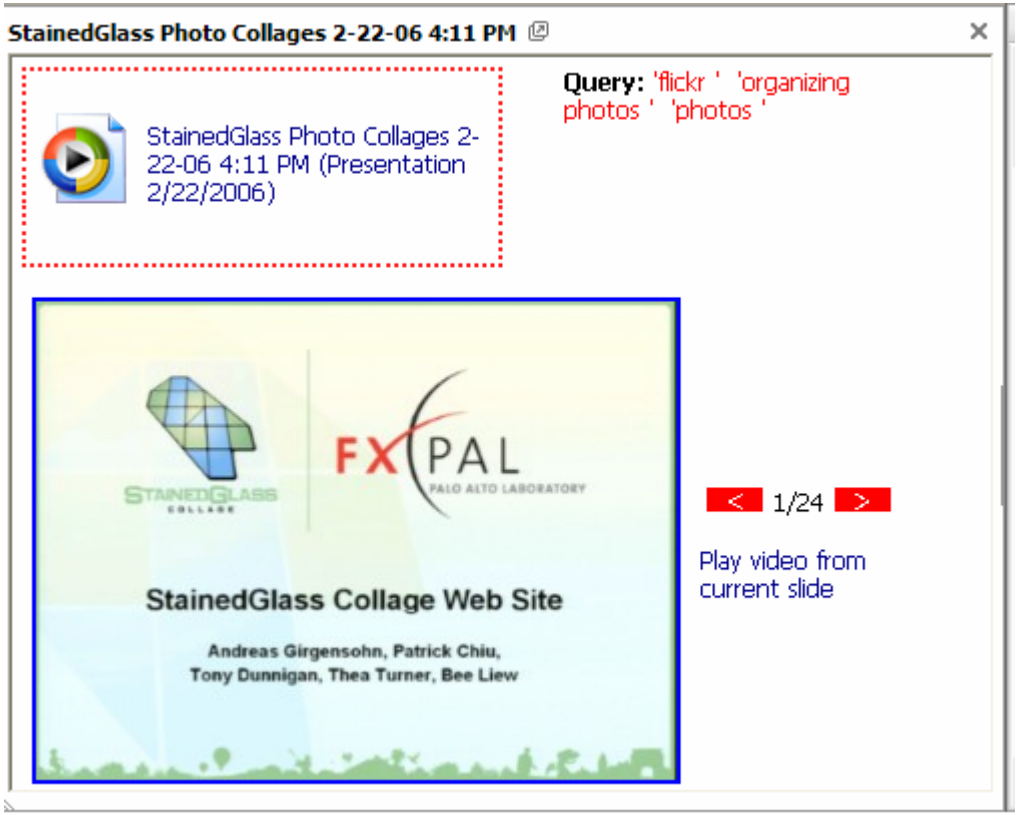
- FXPAL “Corporate Memory”

Query

- FlickR, Organizing Photos, Photos

Result

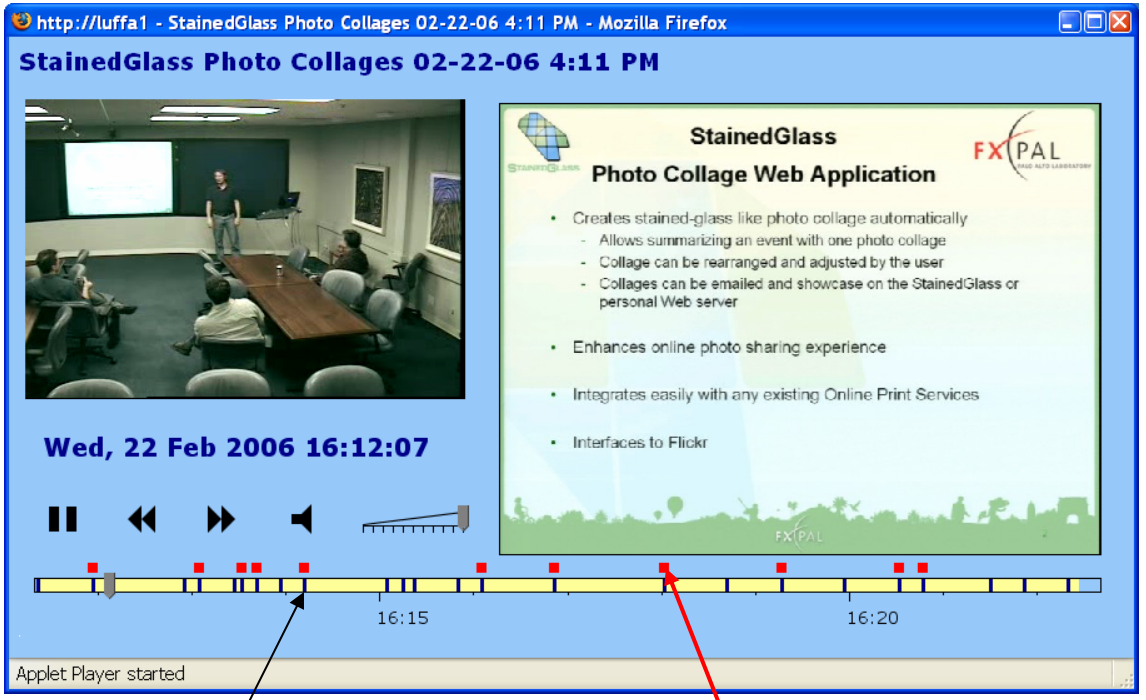
- Slides from presentation
- Click to play lecture video from current slide



Example: Find Segment of Presentation

Play video and view slides

- Skip to next slide
- Skip to next relevant slide



Slides marked by black lines

Relevant Slides marked by red squares

Video Indexing for Stanford Med School

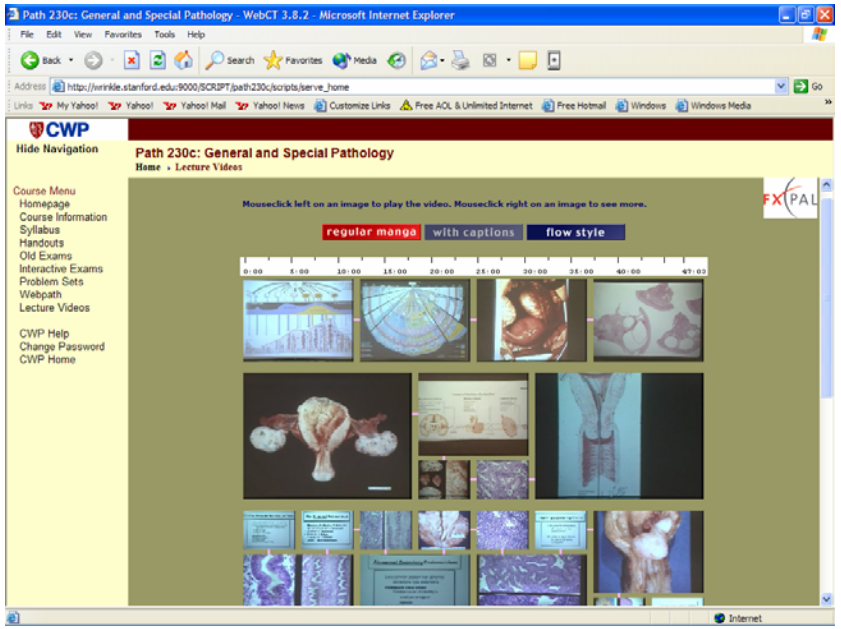
Pathology Class

- Presentation material was mainly images
- No PowerPoint

Create Visual Index into Video

Timeline Slider

- Move cursor along timeline to browse keyframes in video
- Click on keyframe to play
- Good for fast forward/reverse



Manga Summaries

- Visual summary of video
- Click on image to play
- Good for study review

Usage Study

Student Focus Groups

- Primary use of recorded video is to view missed class
- Video sometimes used for review
- Some note-taking while watching recorded lecture but it was too time consuming
- Students want links between video, slides, syllabus, TA handouts, book, other information on Web

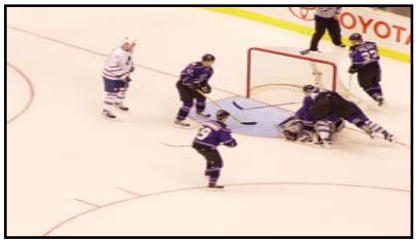
Professor Interview

- Manga and Timeline slider good index for professor since he knows material (slides)
- For students, text index is needed
- Students learn by note-taking and linking information from different sources
- Trial was too short

Video Search – News Programs

Find segments of news on a topic of interest

- Find news story
- Find shots within story



TRECVID

- Sponsored by NIST (National Institute of Standards)
- Data base of 60 hours of news video (ABC, NBC) in 2004 – similar content other years
- Task – user has 15 minutes to find shots relevant to a topic

Example Topics

- *"Find shots of a hockey rink with at least one of the nets fully visible from some point of view"*
- *"Find shots zooming in on the US Capitol dome"*
- *"Find shots of Saddam Hussein"*

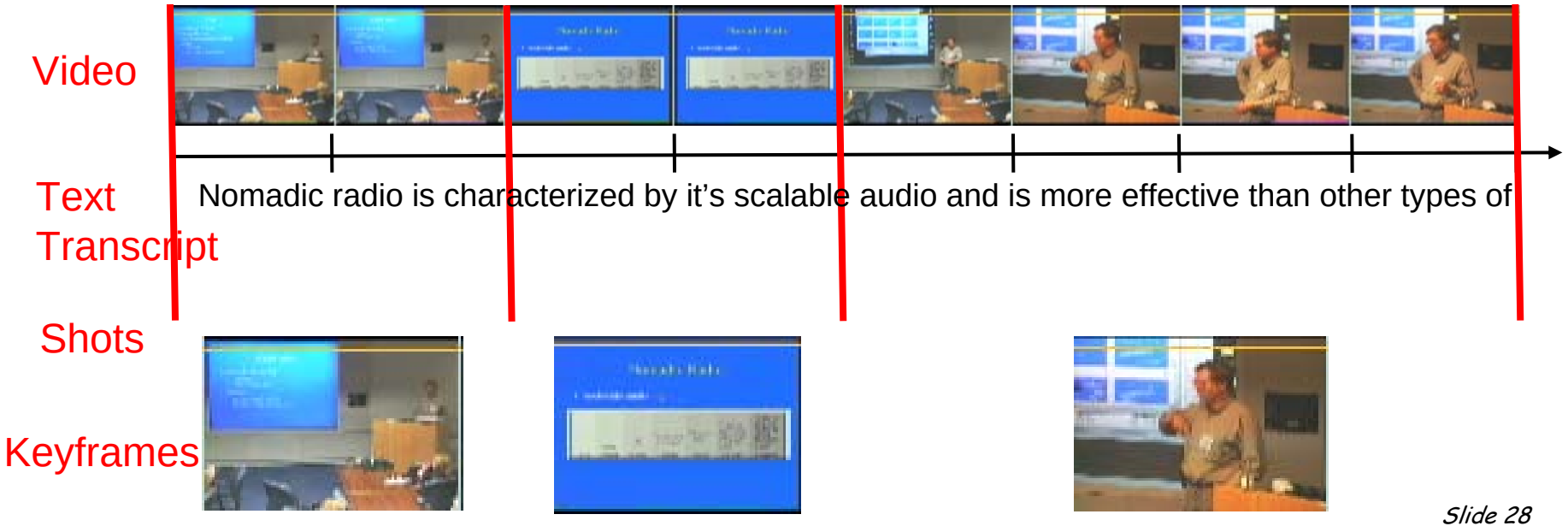
Video Search: Segments of Video

Video
- Sequence of frames (images), typically with audio
- 30 frames/second

Text Transcript of Audio
- Time-correlated with video

Segments of Video
- Shot: Unbroken segment of video from a single camera
- Story (news): Sequence of shots from the same news story

Keyframe
- Representative image from a video segment



Video Search – News Index

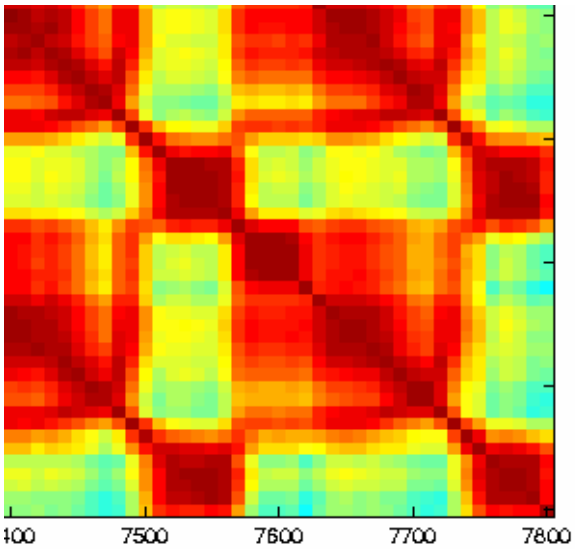
Video Data

- Video - sequence of frames (images)
- Time-aligned text from automatic speech transcription (text)

Pre-processing

- Segment video into shots using image features
  - *Compute pairwise similarity between frames of video*
  - *Similarity is based on image features*
  - *Segment when similarity is low*
- Select a representative keyframe for each shot

- Segment video into stories using text
  - *Compute pairwise similarity between shots of video*
  - *Similarity is based on text associated with shot*
  - *Segment when similarity is low*
- Each story will be composed of one or more shots



Self-similarity matrix for shot and story segmentation

Video Search – News Retrieval

FX PAL

TRECVID task is to find shots relevant to the query

- Use keyword search and image search

Keyword Search

- Retrieve stories relevant to keyword

Image Search

- Retrieve stories with shots relevant to keyword

Merge results of image and keyword search

- Examine shots within the retrieved stories

TRECVID Search

- User enters keywords and/or images for query
- System returns relevant stories
- User explores stories for relevant shots

Story Summary Quads

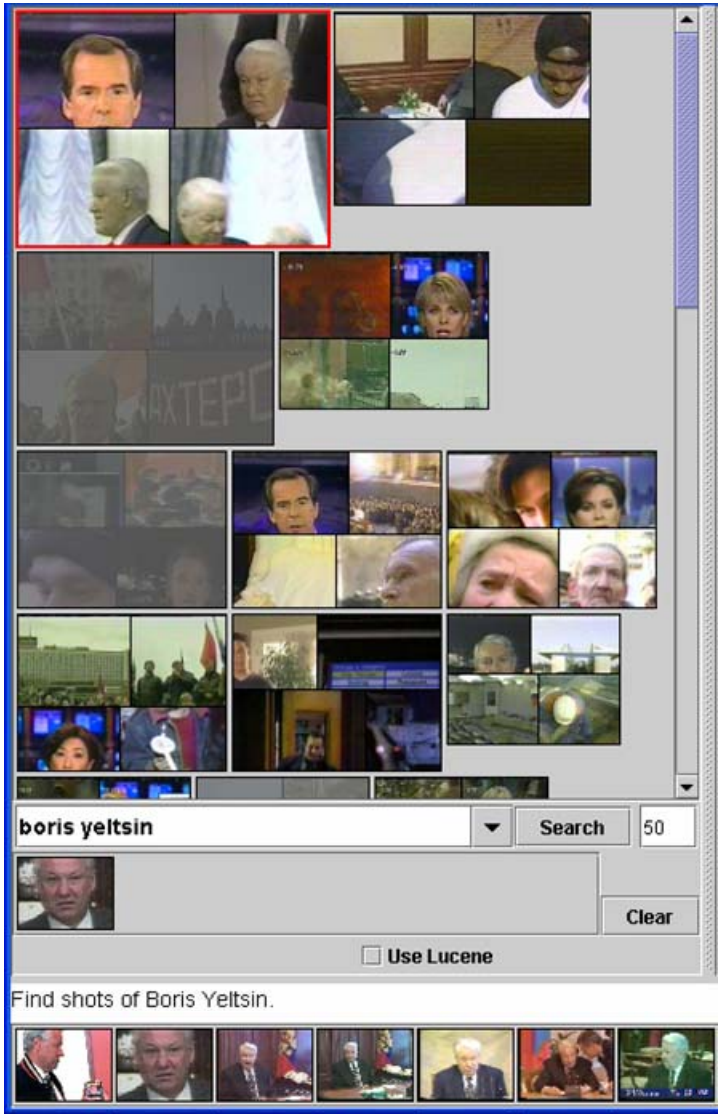
Query-dependent story summary

- Use 4 highest scoring shots
- Allocate space proportional to score

Shot thumbnails

Story thumbnail

Query Area



Search entered as text and example images

- System returns relevant stories

Results displayed in ranked list of story keyframe collages

- Keyframe collages adapt to query

Shading reflects exploration history

Story Browsing Area

Selecting a story expands the video at that point

- Clickable video timeline with relevancy shading
- Clickable story quad timeline
- Overlay on shots marked (non)relevant
- Moused-over icon displays full-size media player area at top and tool-tip shows context of relevancy
- Double click to play video in the media player (if you really have to)


MediaMagic Text Search
Query results area
Media player and zoom area
Relevant shots area
Gray visited overlay
Selected story
Video timeline
Expanded shots area
10/13/1998 ABCa
10/13/1998 ABCa 0:56 [14:51] 0.4
ceased faircloth subsidize reimburse frozen sam donaldson white house
Text query box
Image query box
Excluded overlay
sam donaldson white house
10/13/1998 ABCa 0:02 0.4
sam donaldson white house
Trecvid topic text
Included overlay
Clear
Add related
Select all
Clear
Text search type
Use Lucene
Find shots of Sam Donaldson's face - whole or part, from any angle, but including both eyes. No other people visible with him
11:43
End Question
Trecvid topic images

TRECVID Results

FX PAL

Evaluations based on "Mean Average Precision"

- Recall is percent of relevant shots that are retrieved
  Retrieving all shots give perfect recall
- Precision is the percent of retrieved shots that are relevant
  Retrieving only one relevant shot gives perfect precision
- MAP is a measure of average precision at different recall levels

TRECVID competition results are poor

- On average only 30% of the relevant shots are found correctly

Why Results are Poor

- Keyword is not found in text associated with the video
  Ex: Query is to find people riding bicycles, but no mention of bicycles is made in the text associated with the stories
- Image search not good enough
  General image searches cannot find specific objects like people on bicycles

Multimedia Search Summary

Text Search

- Keywords

Image Search

- Search based on tags (FlickR, FaceBook)
- Search based on surrounding text (Google)
- Content based search
  - *Using image features*
  - *Using faces*

Audio Search

- Search based on metadata (iTunes)
- Content based search (MuscleFish, Foote)

Video Search

- Search based on text (Google/UTube)
- Search based on associated media (Lectures with slides)
- Search based on content (TrecVid News Search)

MediaMagic