Semantic Multi-modal Analysis, Structuring, and Visualization for Candid Personal Interaction Videos

Thesis Proposal: Alexander Haubold

Prof. John R. Kender
(Research Advisor, CS)

Prof. Julia Hirschberg
(Committee, CS)

Prof. Shih-Fu Chang
(Committee, EE)
Overview

1. Introduction: Motivation, Background, Objectives
2. Research Approach
3. Research Progress
   • Structuring Lecture Videos using Visual Contents
   • Structuring Lecture Videos using Textual Contents
   • Analysis of Classroom Presentation Videos
   • Issues in Speaker Clustering
4. Proposed Work
   • High-level Structure Detection and Comparison
   • Speaker Table of Contents
   • Analysis of Discussion Videos
   • Text Indexing
   • User Interface and Tools
   • User Studies
   • Feedback Annotations for Videos (Optional)
5. Conclusion and Schedule
## 1. Background

- Increased video use in universities for recording personal interaction
- Wide range of media: lectures, presentations, talks, meetings, discussions
- Most videos remain unedited but are used by students and instructors
- Analysis of content and structure can provide meaningful indices for browsing and searching
- Analysis must be multi-modal: analysis: video, audio, text

### 1. Introduction

<table>
<thead>
<tr>
<th>2. Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Progress</td>
</tr>
<tr>
<td>4. Proposed Work</td>
</tr>
<tr>
<td>5. Conclusion</td>
</tr>
</tbody>
</table>
Motivation: Interactions in a University

- Lectures:
  (e.g.) 75 minutes * 25 sessions = 31 hours of video

- Monthly student team updates/discussions:
  (e.g.) 100 students (20 teams) * 60 minute discussion * 4 months = 80 hours of video

- Midterm and Final student team presentations:
  (e.g.) 100 students (20 teams) * 15 minute presentation * 2 presentations = 10 hours of video
1. Introduction

Motivation:
Identified shortcomings

• Massive Data: Video material for an interactive 100-student class: 31+80+10=121 hours (≈1.5 Tb)

• No single point of view: instructors and/or students use video for review and/or feedback

• Possibly external parties interested in review: clients

• Content selection varies: What are most important?

• Browsing strategies vary.

• Linear browsing unhelpful: wasted time and effort.
Motivation: Observations

• Videos are rich in multimedia content: image, audio, text, graphics, interactions between people, etc.

• Content summaries should reflect some of these semantics

• Different genres share common features, e.g. all contain spoken text

• Different genres exhibit different emphasis on content, e.g. number of speakers
Related Work

• References from candidacy exam, and additional ones, including:

  • Video: Cornell Lecture Browser [Mukhopadhyay ’97], Video Skimming [Smith ’98], VideoQA [Yang ’03], LSI [Souvannavong ’04], Structuralizing Lecture Video [Dorai ’03], Visual Semantics [Natsev ’04], Visual Concepts [Kender ’05], Segmenting People [Lee ’02]

  • Audio: Speaker Segmentation [Chen ’98], Audio Features for Scene Segmentation [Liu ’98], SCAN – Speech Archives [Whittaker ’99], Video Mail [Young ’97], SMaRT Meeting Room [Waibel ’03], Speech Recognition Experiments [Witbrock ’97]

  • Text: LSA [Landauer ’98], Lecture Video Retrieval [Fujii ’03], Lecture Audio Data [Glass ’04], Video Segmentation using Text [Lin ’04], Salient Segments [Ponceleon ’01], WordNet [Fellbaum ’98], Redundant Words [Yang ’96]

  • A/V/T: Video Retrieval using Speech and Image [Hauptmann ’03], Audio-Visual Structure in Film [Sundaram ’00]

  • UI: UI Issues for Browsing Video [Lee ’99], The Eyes Have It [Shneiderman ’96], Browsing Digital Video [Li ’00], Keyframe Indexing [Girgensohn ’01], Intelligent UI [Tang ’06], Semantic Browser [Altman ’02]
Research Objectives

- Establish multi-modal indices, including video, audio, and text, for three candid interaction video genres
- Determine structure of content and interaction in videos to build an analog to a book’s TOC
- Construct tools for segmentation and visualization
  - Identify and build upon commonalities
  - Identify and customize upon uniqueness
- Introduce a structure comparison tool to highlight interesting deviations in a set of videos
2. Research Approach

• Specify semantic characteristics for three genres
  – Common
  – Unique to each genre
• Process videos:
  – Segmentation/clustering of audio and video
  – Text extraction and augmentation
• Visualize information in a browser
# Three Genres of Personal Interaction Videos

- Lectures, Presentations, Discussions

<table>
<thead>
<tr>
<th>Type</th>
<th>Speaker</th>
<th># Speakers</th>
<th>Audience</th>
<th># Audience</th>
<th>Structure</th>
<th>External Material</th>
<th>Recording Operator</th>
<th>Camera/Microphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture</td>
<td>Expert</td>
<td>1</td>
<td>Novices</td>
<td>1</td>
<td>Explicit</td>
<td>Written content</td>
<td>Trained</td>
<td>Fix/fixed</td>
</tr>
<tr>
<td>Classroom Presentation</td>
<td>Novices</td>
<td>N</td>
<td>Expert</td>
<td>1</td>
<td>Implicit</td>
<td>Structured outline</td>
<td>Untrained/none</td>
<td>Fix/mobile</td>
</tr>
<tr>
<td>Discussion</td>
<td>Novices</td>
<td>N</td>
<td>Novices/expert(s)</td>
<td>N</td>
<td>None</td>
<td>None</td>
<td>Untrained/nervous</td>
<td>Mobile/mobile</td>
</tr>
</tbody>
</table>

1. Introduction
2. Approach
2.1 Three Genres
3. Progress
4. Proposed Work
5. Conclusion
Common Tools for Analysis and Visualization

• Audio contains cues about interaction breaks
• Video contains keyframes for visual summaries
• Text from speech lends cues into text semantics
• Structure is formed from combination of modalities
Common Tools for Analysis and Visualization

• Each modality requires separate consideration for
  • Analysis, and
  • Visualization
Common Tools for Analysis and Visualization

• We have identified several commonalities
• Work on these has been completed
Common Tools for Analysis and Visualization

- We propose additional common tools for analysis and visualization
Common Tools for Analysis and Visualization

- Genre-specific “modules” are required to capture their characteristics

Solid = completed, Dashed = proposed
Common Tools for Analysis and Visualization

• Modules are integrated with common tools

Solid = completed, Dashed = proposed
Common Tools: Audio
WHO is speaking?

- Determine interaction and recurrence between speakers (topology)
- Speaker segmentation:
  - BIC based on MFCC features
  - Largest difference between 2 speaker MFCC distributions
- Speaker clustering:
  - Symmetric KL based on MFCC
  - Pair-wise comparison of speaker feature distributions
Common Tools: Video
WHERE is it taking place?

• Most interaction videos lack “action”
• Analog to video player: Keyframe player for fast browsing/skimming
  • Keyframes sufficient as summaries
  • Genre-specific consideration of visual features
Common Tools: Text
WHAT is being said

• Provides text semantics
• Word/phrase indices within and across videos
  • No training on speech/language models (infeasible)
  • ASR: high (75%) error rates
  • Filter meaningful terms using external dictionary (dependent on genre)
Common Tools: Structure
Tying it all together

• Modalities gives clues about structure of video
• Establish hierarchical Table of Contents
  • Combine structures of separate modalities
  • Rank segments among modalities by overlapping
  • Build longest-to-shortest hierarchy
Genre-specific tools: Lectures

• Video segments are teaching units that relate to the textbook

(Activity:)

• Video:
  • Segmentation between periods of inactivity (proprietary)
  • Pixel content clusters (PCC) establish teaching units
  • Environmental clusters for domain-specific location
  • Video content index from representative frame from PCC
  • Video scene topology from tracking content over time

(Text:)

• Related course textbook indices used as external filter
• Lecture clustering through recurring key phrase

(Structure:)

- Textbook indices
- Video segment clustering
- Environmental clustering
- Pixel content clusters
- Video scene topology

Diagram of analysis and core modules.
Genre-specific tools: Presentations

- Presentations are units of common content addressed by several students

<table>
<thead>
<tr>
<th>Audio:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Annotation of speaker segments with face shot from video</td>
</tr>
<tr>
<td>• Visual speaker index</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Segmentation on significant content and motion changes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Electronic slides used as external filter</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(Structure:)</th>
</tr>
</thead>
</table>

---

![Diagram of genre-specific tools](image_url)
Genre-specific tools: Discussions

• Discussions cover several topics of interest, i.e. points on an agenda

<table>
<thead>
<tr>
<th>Audio:</th>
<th>Text:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Annotation of speaker segments with face shot from video</td>
<td>• Meeting agenda or reports on project used as external filter</td>
</tr>
<tr>
<td>• Visual speaker index</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Segmentation on significant motion changes</td>
</tr>
<tr>
<td>• Environmental clusters (EC) for domain-specific location</td>
</tr>
<tr>
<td>• Video content index from representative frame from EC</td>
</tr>
</tbody>
</table>

(Structure:)
3. Research Progress

• Structuring Lecture Videos using visual contents (*ICME 2003*)
• Structuring Lecture Videos using textual contents (*MCBAR 2004*)
• Segmentation and Augmentation for Classroom Presentation Videos (*MM 2005*)
• Research on Accommodating Sample Size Effects in symmetric KL in Speaker Clustering (*MM 2006 submitted*)
3.1 Structuring Lecture Videos using Visual Contents (ICME 2003)

- Video-taped courses made available to students with “keyframe” index
- "Keyframe" = snapshot of video at points of substantial change every 20-25 seconds
- Typical course length: 75 min per lecture, 26 lectures per semester = 32.5 hr of video data (3.5 megaframes!) => 5000 snapshots
- Need a more compact, content-directed indexing method for keyframes
Process Overview


5. Visualization and Interface ← 4. Clustering by Content
Environment Clusters
“Media Types”

- Snapshots belong to six Media Types
- Apply decision tree classifier

<table>
<thead>
<tr>
<th>Color</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>Board</td>
</tr>
<tr>
<td>Blue</td>
<td>Podium/Instructor</td>
</tr>
<tr>
<td>Red</td>
<td>Computer</td>
</tr>
<tr>
<td>Yellow</td>
<td>Hand-drawn Sheets</td>
</tr>
<tr>
<td>Orange</td>
<td>Printed Media (Illustrations)</td>
</tr>
<tr>
<td>Cyan</td>
<td>Class</td>
</tr>
</tbody>
</table>
Clustering by Content Matching

- Snapshot content evolves slowly
- Hand-drawn slides grow monotonically
- Blackboard panels vary dynamically
- Apply content pixel filter
- Comparisons between sub-windows of keyframes
- Matching by minimizing content pixel differences, sub-window translation vector
Interface and Visualization
Interface and Visualization: Topological View

- Temporal topic model with relative time scale
- User study confirms that perception is topological
- Topics represented with media icons
- Interrupted topics reunited by tapering lines
- Structural significance in nesting of topics
Interface and Visualization:
Keyframe view

• Abstract topics in context through keyframes
• Each media icon linked to original keyframes
• Fast browsing using mouse-over pop-ups or keyframe player
• User study confirms increase in browsing speed
3.2 Structuring Lecture Videos using Textual Contents (MCBAR 2004)

• Automatic Speech Recognition (ASR) Transcripts typically used for searching, categorization of video databases
• Lecture videos = dozens of contextually connected entities
• Typical course: 10 to 30 lectures (70 or 120 min);
  Typical lecture: 5k – 14k words ≈ 150k words
• Need indices across lectures, courses
• Extract and display structure of entire course using key words/phrases
Process Overview

... aum sell and it is its structure doesn’t provide a way to find something like a binary tree provides a way of looking for 27 and by treat it is given a pointer to the router the treehouse where it ought and emulate ...

Imperfect Transcript

Interactive Visualization

Lecture Audio

Textbook Index or Manual Index

1. Introduction
2. Approach
3. Progress
3.2 Lecture Videos: Textual
4. Proposed Work
5. Conclusion

Index Words and Phrases

Word Pairs

- find binary
- find structure
- find tree
- binary tree
- binary pointer
- binary structure
- tree pointer
- tree structure

Imperfect Transcript

Big Endian
Binary
search
tree
...
Pointers
Transcript Generation

• ASR transcript from IBM® ViaVoice®
• Highly compressed lecture video
  • Poor audio quality
  • Word Error Rate of 75%
• Training: little (3%) overall improvement
  • Number of unique index phrases ≈ same
  • But: difference in identified index phrases 20%
  • Best to combine trained and untrained results
Target Corpus

• Lectures: rich in subject-specific terms
• Define:
  • *Theme phrases*: General tenor for contents of course
    • in many (> ¼) transcripts
  • *Topic phrase*: Highlight specific topics for lectures
    • in few (< ¼) transcripts
  • *Illustration phrases*: unique terms for examples
    • Hard to identify in highly imperfect transcripts
Filtering Index Phrases

- Structured approach
- Use corpus of expected phrases: index of course textbook
  - Capture key phrases of length 1-3
  - Rarely longer; index reflects likelihood
  - (1) Collapse indentation hierarchy
  - (2) Remove stop words in beginning & end of each line
  - (3) Stem

<table>
<thead>
<tr>
<th>amortized analysis</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>accounting method of</td>
<td>(1,2,3)</td>
</tr>
<tr>
<td>aggregate analysis of</td>
<td>(1,2)</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>call by value</td>
<td></td>
</tr>
</tbody>
</table>
Filtering Word Pairs

• Unstructured approach
  • Address: speech in lecture fragmented
• Use textbook index to filter keywords
  • Remove structure from index: use only words
  • Word pairs in transcript = index words separated by ≤ 10 words

<table>
<thead>
<tr>
<th>multiple instruction</th>
<th>call structural</th>
</tr>
</thead>
<tbody>
<tr>
<td>multiple operation</td>
<td>call hazard</td>
</tr>
<tr>
<td>multiple very</td>
<td>call instruction</td>
</tr>
<tr>
<td>multiple word</td>
<td>call compaction</td>
</tr>
<tr>
<td>multiple processor</td>
<td>call step</td>
</tr>
</tbody>
</table>
Results of Filtering

• Analysis of 273 transcripts (11 courses)

• Index Phrases
  • Unique per textbook indices: 253-4701
  • Unique per transcript: 8-98
  • Occurrence: 1 (35-50%), 5-50 (20-30%)
  • Unique per course: 40-347

• Word Pairs
  • ≈ 10 times more than index phrases
  • Less meaningful for summaries
Interface: Parallels to a Camera

• 3 visualization techniques (2 discussed here)
• Share 3 freely variable parameters:
  • Zoom: specificity of phrases
    • Occurrence of phrase across transcript
    • Range: Topic-specific to entirely thematic
  • Focus: emphasis of phrases
    • Range: 1 – N (lowest – highest occurrence)
  • Contrast: length of phrases
    • Range: 1 – K (K usually 3)
2. Approach

1. Introduction

3. Progress

3.2 Lecture Videos: Textual

4. Proposed Work

5. Conclusion

Interface: Transcript Index Map

- Index phrases mapped to transcript
- Equivalent to textbook index
  - But: order by occurrence (highest near top)
  - Color coded (red → yellow = high → low occ.)
- Cross-reference terms among transcripts
  - Longer blobs for repeating phrases
- Greedy population of space near top
2. Approach

1. Introduction

Interface: Chapter Transcript Match

- Transcripts mapped to textbook chapters
- Rows=transcripts, Columns=chapters
- Match score based on occurrences of terms between transcript and chapter
- Best performance (70% accuracy) when using combination of Index Phrases and Word Pairs

| 01:transcripts | 02:transcripts | 03:transcripts | 04:transcripts | ALL | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 |
| 01:transcripts | 02:transcripts | 03:transcripts | 04:transcripts | ALL | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 |
| 01:transcripts | 02:transcripts | 03:transcripts | 04:transcripts | ALL | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 |

3. Progress

3.2 Lecture Videos: Textual

4. Proposed Work

5. Conclusion
3.3 Segmentation and Augmentation for Classroom Presentation Videos (MM 2005)

<table>
<thead>
<tr>
<th>Videos of student team presentations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 semester ≈ 160 students, 30 teams, 8 hours of video for midterm presentations</td>
</tr>
<tr>
<td>How to best review?</td>
</tr>
<tr>
<td>Need automatic index for videos</td>
</tr>
<tr>
<td>Need visual browser for searching</td>
</tr>
</tbody>
</table>
Characteristics

• Multiple speakers: ≈ 5 / team, ≈ 20 / hour
• Not professionally recorded or edited
• Lighting conditions vary
• Long shots without distinct visual cuts
• Audio quality varies (handling of microphone)
• But: known structure of thematic sections
Process Overview

1. Introduction
2. Approach
3. Progress
3.3 Presentation Videos
4. Proposed Work
5. Conclusion
Segmentation (Audio)

<table>
<thead>
<tr>
<th>Audio:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Identify audio segments for each student</td>
</tr>
<tr>
<td>• MFCCs for representing features of speech</td>
</tr>
<tr>
<td>• Bayesian Information Criterion detects speaker changes</td>
</tr>
<tr>
<td>• Results encouraging, even for varying audio quality</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video: Boundaries from non-overlapping sources:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Presentation slide changes</td>
</tr>
<tr>
<td>• Not all presentations have slides</td>
</tr>
<tr>
<td>• Speaker gesture changes</td>
</tr>
<tr>
<td>• Long-term change in speaker pose</td>
</tr>
<tr>
<td>• Reconfiguration of speaker position</td>
</tr>
<tr>
<td>• Amount of gesture</td>
</tr>
</tbody>
</table>
Segmentation
Combined Audio/Visual

• Combination of audio and video cues results in more natural segmentation
  – Not every speaker change is accompanied by visual change, and vice versa
  – Presentation Unit: Union of A/V change
Text Augmentation

• Follows our previous work (MCBAR 2004)
• No language/model training (would require 160 / semester)
• Apply 2 filters
  • Theme phrases: manually assembled list
  • Phrases / titles of required sections
  • Topic phrases: presentation slides (if available)
    • Appear in presentation AND transcript
• Theme Phrases:

<table>
<thead>
<tr>
<th>“alternative solutions”</th>
<th>“objective tree”</th>
<th>“background”</th>
<th>“functional”</th>
<th>“prototype”</th>
</tr>
</thead>
<tbody>
<tr>
<td>“continuity plan”</td>
<td>“problem statement”</td>
<td>“chart”</td>
<td>“future”</td>
<td>“requirements”</td>
</tr>
<tr>
<td>“design constraints”</td>
<td>“project goals”</td>
<td>“constraints”</td>
<td>“goal”</td>
<td>“schedule”</td>
</tr>
<tr>
<td>“functional requirements”</td>
<td>“tasks performed”</td>
<td>“continuity”</td>
<td>“implementation”</td>
<td>“solutions”</td>
</tr>
<tr>
<td>“future directions”</td>
<td>“team process”</td>
<td>“deliverables”</td>
<td>“limitations”</td>
<td>“statement”</td>
</tr>
<tr>
<td>“gantt chart”</td>
<td>“team development”</td>
<td>“demo”</td>
<td>“objective”</td>
<td>“tasks”</td>
</tr>
</tbody>
</table>
Interface

- List of Videos
- Video Playback
- Zoomable Summary
  - Thumbnails
  - Timeline
- Audio, video tracks
- Text tracks
3. Progress

3.3 Presentation Videos

- Portrait notebook-style not well received
- Re-modeled to horizontal continuous timeline
Interface: Text Graph

• Zoomable interface distributes text
  • 10 minutes
  • Deeply nested text
  • 1.5 minutes
  • More precise browsing
User Study

• 176 students, mostly appearing in videos
• Questions answered using UI

<table>
<thead>
<tr>
<th>Find your appearance during presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find beginning of your team’s presentation</td>
</tr>
<tr>
<td>Find your team’s discussion on topic X</td>
</tr>
<tr>
<td>Find presentation Y (Y of different team &amp; class)</td>
</tr>
<tr>
<td>Summarize segment using only text</td>
</tr>
</tbody>
</table>

• ½ students: summaries + video playback
• ½ students: only summaries
User Study: Results

• Video + Summaries vs. Summaries only
  • Overall same accuracy
  • 20% less time spent without video
  • But: no comparison to linear search (VCR)

• System
  • External structure of contents important
    • Apply and visualize in browser
  • Zoomable text requires ranking (structure)

• User
  • Thumbnails good: focus on task
  • Video bad: easily sidetracked
3.4 Research on Speaker Clustering (submitted to MM 2006)

- Popular approach: comparison using symmetric KL
  \[ KL2(A, B) = C(A, B) + M(A, B) \]
  \[ C(A, B) = \frac{1}{2} tr(\sigma_A^{-1}\sigma_B) + tr(\sigma_B^{-1}\sigma_A) - d \]
  \[ M(A, B) = (\mu_A - \mu_B)(\sigma_A^{-1} + \sigma_B^{-1})(\mu_A - \mu_B)^T \]
  (\(\sigma\) is covariance matrix, \(\mu\) is mean vector, \(d\) is dimensionality of feature set, e.g. MFCC)

- KL2 based on comparison between Gaussian distributions

- Observation: performance degradation for comparisons between differently-sized feature sets (speaker segments)

- Problem occurs for speech segments of length < 30-60 sec
Sample Data

- Example comparisons between differently-size feature sets
- The longer two segments of the same speaker, the better they cluster

<table>
<thead>
<tr>
<th>Length(MFCCa)</th>
<th>Length(MFCCb)</th>
<th>KL2</th>
<th>KL2 w/o mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20</td>
<td>55.28</td>
<td>39.034</td>
</tr>
<tr>
<td>20</td>
<td>2000</td>
<td>30.688</td>
<td>22.135</td>
</tr>
<tr>
<td>20</td>
<td>4000</td>
<td>29.657</td>
<td>24.308</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>4.1154</td>
<td>2.4293</td>
</tr>
<tr>
<td>100</td>
<td>2000</td>
<td>2.4607</td>
<td>1.6848</td>
</tr>
<tr>
<td>100</td>
<td>4000</td>
<td>1.9112</td>
<td>1.407</td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>1.7362</td>
<td>0.9844</td>
</tr>
<tr>
<td>500</td>
<td>2000</td>
<td>0.21889</td>
<td>0.15538</td>
</tr>
<tr>
<td>500</td>
<td>4000</td>
<td>0.36733</td>
<td>0.28415</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>0.54498</td>
<td>0.35734</td>
</tr>
<tr>
<td>1000</td>
<td>4000</td>
<td>0.23357</td>
<td>0.13032</td>
</tr>
<tr>
<td>2000</td>
<td>2000</td>
<td>0.066197</td>
<td>0.044169</td>
</tr>
<tr>
<td>2000</td>
<td>4000</td>
<td>0.066557</td>
<td>0.038734</td>
</tr>
<tr>
<td>3000</td>
<td>3000</td>
<td>0.009779</td>
<td>0.006934</td>
</tr>
<tr>
<td>4000</td>
<td>4000</td>
<td>1.42E-05</td>
<td>9.40E-06</td>
</tr>
</tbody>
</table>
Simulation of KL2

- Simulated KL2:
  - For large number of random distributions (size 20-1M)
  - For real speech data (sampled random segments)
  - Problem apparent in regions of small feature sets
Empirical Solution

• Introduce factor by which KL2 is offset depending on length of segments
• Use simulation results as matrix of look-up values

\[ KL2' = \frac{KL2(A, B)}{KL2_{sim}(|A|, |B|)} \]

• KL2’ is adjusted symmetric KL distance
Results

• Evaluated approach on clustering speakers in presentations
• Many speech segments > 1 minute, some 20-40 seconds
• Clustering with KL2 does not correctly cluster short speech
• Clustering with KL2’ corrects values, while not adversely affecting other clusters

KL2 without scaling

KL2 with scaling
Visualization Improvements

• With speaker clustering, UI can be improved to include speaker topology
• Parallel to content topology for lecture video
• Ideally: extract faces from video as speaker labels
3.5 Other Related Completed Work

• Visualization for Periodic Population Movement between Distinct Localities (*InfoVis 2003*)
  • 2D map highlighting substantial migration over time
• Lighting Control using Pressure-Sensitive Touchpads (*Maker Faire 2006*)
  • Touchpad (mouse) used for controlling multi-colored LED lighting
• Semantic Multimedia Retrieval Using Lexical Query Expansion and Model-based Re-ranking (*ICME 2006*)
  • Visual concept evaluation for textual query expansion
• Improving Semantic Multimedia Retrieval Using Statistical Model-based Re-ranking (*MM 2006 submitted*)
  • Statistical mapping of visual concepts to text for improving text queries
4. Proposed Work

- In order of importance:
  - High-level Structure Detection
  - Video Structure Comparison
  - Speaker Table of Contents
  - Analysis of Discussion Video and Application of Common Approaches
  - Text Indexing
  - User Interface and Tools
  - User Studies
  - Feedback Annotations for Videos (Optional)
4.1 High-level Structure Detection

• **Hypothesis:** Humans perceive video in semantic units and structure

• Books are organized in a similar fashion (chapters, sections)

• Especially for content-rich video, determining structure is necessary for enhanced browsing

• **To Do:** Determine multimedia hierarchy for three genres

<table>
<thead>
<tr>
<th></th>
<th>Book</th>
<th>Chapter</th>
<th>Section</th>
<th>Paragraph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture Video</td>
<td>All lectures for a semester</td>
<td>One lecture</td>
<td>Teaching units: topics of similar written content</td>
<td>One keyframe</td>
</tr>
<tr>
<td>Student Presentation Video</td>
<td>All presentations for a class</td>
<td>One presentation</td>
<td>Material discussed by one speaker; Q&amp;A session</td>
<td>Electronic Slide (if available)</td>
</tr>
<tr>
<td>Discussion Video</td>
<td>All discussions for a team</td>
<td>One discussion</td>
<td>Dialogue between N partakers on a topic</td>
<td>Speaker</td>
</tr>
</tbody>
</table>

Potential parallels to video structure
4.2 Video Structure Comparison

• **Hypothesis:** Genres have characteristic style

• High-level comparisons possible to
  - Cluster videos
  - Find interesting deviations

• **To Do:** Determine grammar by which structure can be compared
4.3 Speaker Table of Contents

• **Hypothesis:** Visual speaker index powerful TOC
• Video summaries can be browsed quickly using keyframes
• Audio can be browsed quickly using text from speech
• Speaker identity can be extracted from speech (voice sample), but is still serial in nature
• Visual speaker indices allow for fast browsing of speakers in a video
• **To Do:** Extraction of talking heads from video
4.4 Analysis of Discussion Video and Application of Common Approaches

• **Hypothesis:** Discussion videos are least structured, but there exists some structure

• Speakers are likely an indicator of people interactions
  • Is the discussion overshadowed by many interruptions?

• **To Do:** Discover structure using tools built for stronger genres
4.5 Text Indexing

• **Hypothesis:** Text is rich cue, but can be overwhelming

• After filtering, some phrases remain but context is lost

• Deep analysis using tools like WordNet can help in finding meaning for group of seemingly random words

• Is it possible to build short meaningful summaries from filtered text?

• **To Do:** Explore semantic network, like WordNet for sense disambiguation and determination of themes
4.6 User Interface and Tools

**Hypothesis:** Analysis and Visualization tightly linked: they drive each other

- Further integration of analysis and visualization tools required to build tool for
  - Use in the classroom environment
  - Evaluation of research

**To Do:** Improve and introduce new analysis and visualization tools
4.7 User Studies

• **Hypothesis:** Eventually, the viewer decides over utility

• 4 user studies completed
  • Evaluation of topological lecture browsing
  • 3-semester evaluation of improvements of presentation video UI

• **To Do:** Continuing student evaluation of tool ($\approx 160 / \text{semester}$) with targeted search and browsing tasks

• **To Do:** integration in classroom environment

• **To Do:** IRB approval for future user studies
4.8 Feedback Annotations for Videos (Optional)

• **Hypothesis:** Students benefit from instructor feedback
  • Improve presentation skills given good/bad examples

• Qualitative annotation of presentation videos by instructor a possible solution

• **To Do:** Build UI tool by which videos are annotated and information is presented to viewers
5. Conclusion

• Schedule
• Conclusion
5.1 Schedule
Dependency Graph

Audio Structure (4.3) -> User Studies (4.7) -> Text Indexing (4.5)

Structure (4.1, 4.2) -> Discussion Videos (4.4) -> Lecture Videos (4.6) -> User Interface (4.6)

Annotations (4.8)
## 5.1 Schedule

### Calendar

<table>
<thead>
<tr>
<th>Task</th>
<th>Subtasks</th>
<th>Timeframe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Structure</td>
<td>Speaker clustering, modified GMM</td>
<td>Spring 2006</td>
</tr>
<tr>
<td>(Hypothesis 4.3)</td>
<td>Speaker segmentation at higher temporal precision</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Visual Speaker Indices (Face Detection)</td>
<td></td>
</tr>
<tr>
<td>Text Indexing</td>
<td>Segmentation by key words/phrases</td>
<td>Summer 2006</td>
</tr>
<tr>
<td>(Hypothesis 4.5)</td>
<td>ASR Transcript / Audio synchronization</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Text querying / search</td>
<td></td>
</tr>
<tr>
<td>User Interface</td>
<td>Streaming video</td>
<td>Summer 2006</td>
</tr>
<tr>
<td>(Hypothesis 4.6)</td>
<td>Keyframe player</td>
<td></td>
</tr>
<tr>
<td>Annotations</td>
<td>Text annotation of videos for feedback, comments</td>
<td>Summer 2006 (Optional)</td>
</tr>
<tr>
<td>(Hypothesis 4.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion Videos</td>
<td>Visual segmentation</td>
<td>Fall 2006 (suspended from course for academic year 2005-06)</td>
</tr>
<tr>
<td>(Hypothesis 4.4)</td>
<td>Audio segmentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Transcript analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Application of approaches from two other genres)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Merge visual segmentation with transcript clustering</td>
<td></td>
</tr>
<tr>
<td>Lecture Videos</td>
<td>Implement into video browser</td>
<td>Spring 2007</td>
</tr>
<tr>
<td>(Hypothesis 4.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>User studies</td>
<td>Unsupervised usage logging</td>
<td>Spring 2006 – Fall 2007</td>
</tr>
<tr>
<td>(Hypothesis 4.7)</td>
<td>Supervised experiments (IRB)</td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td>Tool and methods for video comparison using structural cues</td>
<td>Fall 2007</td>
</tr>
<tr>
<td>(Hypotheses 4.1, 4.2)</td>
<td>Highlight structural differences between videos of same genre</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster using structure (video similarity)</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

• Integrated approach to analysis of selected video genres

• Novel visualization techniques to engage in more effective browsing of video content

• Introduction of a book-like TOC for content-rich video

• Introduction of structure comparison between videos based on multi-modal analysis

• Common Core modules of analysis and visualization apply to all three genres

• Genre-specific modules to capture characteristics
Thank you!