Indexing and Browsing Unstructured Videos using Visual, Audio, Textual, and Facial Cues

Thesis Defense: Alexander Haubold
## Overview

<table>
<thead>
<tr>
<th>Section</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Introduction</td>
</tr>
<tr>
<td>2.</td>
<td>VAST MM: An orientation</td>
</tr>
</tbody>
</table>
| 3.      | Lecture Videos  
  • Visual Indices  
  • Textual Indices  
  • Task-based Experiments and Evaluation  
  • Course-based Experiments and Evaluation |
| 4.      | Presentation Videos  
  • Introduction  
  • Visual Indices  
  • Audio Indices  
  • Textual Indices  
  • Task-based Experiments and Evaluation |
| 5.      | Common Indices  
  • Audio-Text Alignment  
  • Keyword and Key phrase UI  
  • Annotations  
  • Taxonomy Browser |
| 6.      | Characteristics of Video Browsing |
| 7.      | Conclusion and Future Work |
1. Background

- Increased use of video in universities
- Media: lectures, presentations, talks, meetings, discussions
- Mostly unedited (raw, unstructured) but used by students and instructors
- Storage easy (and very cheap), but organization, dissemination difficult, in most cases not performed
Motivation: Interactions in a University

• Lectures:
  (e.g.) 75 minutes * 25 sessions = 31 hours of video / course / semester

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

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

- Video Segmentation: Cornell lecture browser [Mukhopadhyay ’97], Video skimming [Smith ’98], VideoQA [Yang ’03], LSI [Souvannavong ’04], Structuralizing lecture video [Dorai ’03]

- Audio Segmentation: Text-from-speech topics [Fujii ’03], Video classification [Liu ’98], Speaker segmentation [Chen ’98]

- Text-from-speech: SMaRT meeting room [Waibel ’03], SCAN – speech archives [Whittaker ’99], Video mail [Young ’97], Speech recognition experiments [Witbrock ’97], Language models for lectures [Glass ’04]

- Text Segmentation: Redundant words [Yang ’96], Video segmentation using text [Lin ’04], Salient segments [Ponceleon ’01], LSI [Landauer ’98]

- User Interfaces: 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]

- Structure: Conceptualization [Natsev ’04], Video Retrieval using Speech and Image [Hauptmann ’03], Audio-Visual Structure in Film [Sundaram ’00]
2. VAST MM Orientation

- What is it?
- What video content is indexed?
- Tested on large database of videos
  - ~ 1,000 hours of video
- History of its development
  - 2005 - 2008
What is it? What makes it unique?

- VAST MM = *Video Audio Structure Text MultiMedia*
- Multimedia indexer and browser:
  - Large libraries (> 100 hours of video)
  - Unstructured videos
  - Multi-modal cues (visual, audio, text)
  - Vary displayed information interactively
  - Extensive User Studies
Content Indexing / Storage

2. VAST MM Orientation
1. Introduction / Background
2. VAST MM Orientation
3. Lecture Videos
4. Presentation Videos
5. Common Indices
6. Video Browsing Characteristic
7. Conclusion
## Our Video Database

<table>
<thead>
<tr>
<th>Year</th>
<th>Midterm Presentations</th>
<th>Final Presentations</th>
<th>Student Introductions</th>
<th>Team Meetings</th>
<th>Team Interactions</th>
<th>Student Interviews</th>
<th>Lectures</th>
<th>Events</th>
<th>Talks</th>
<th>Num</th>
<th>Time (Hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Num</td>
<td>Time</td>
<td>Num</td>
<td>Time</td>
<td>Num</td>
<td>Time</td>
<td>Num</td>
<td>Time</td>
<td>Num</td>
<td>Time</td>
<td>Num</td>
</tr>
<tr>
<td>1996</td>
<td>1</td>
<td>1067</td>
<td>1</td>
<td>774</td>
<td>1</td>
<td>1854</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>7</td>
<td>17980</td>
<td>13</td>
<td>35398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>10</td>
<td>31571</td>
<td>20</td>
<td>64643</td>
<td>9</td>
<td>20004</td>
<td>4</td>
<td>12432</td>
<td>19</td>
<td>50253</td>
<td>5</td>
</tr>
<tr>
<td>2003</td>
<td>17</td>
<td>49083</td>
<td>22</td>
<td>69304</td>
<td>5</td>
<td>8556</td>
<td>32</td>
<td>87133</td>
<td>65</td>
<td>167855</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>19</td>
<td>64693</td>
<td>23</td>
<td>75946</td>
<td>25</td>
<td>71283</td>
<td>38</td>
<td>122705</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>16</td>
<td>48405</td>
<td>15</td>
<td>50144</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>16</td>
<td>45734</td>
<td>16</td>
<td>41486</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>12</td>
<td>39187</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>411</td>
<td>330</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Videos - Columbia Video Network**

<table>
<thead>
<tr>
<th>Year</th>
<th>Num</th>
<th>Time (Hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>39</td>
<td>190198</td>
</tr>
<tr>
<td>2003</td>
<td>107</td>
<td>476629</td>
</tr>
<tr>
<td>2004</td>
<td>253</td>
<td>1327296</td>
</tr>
<tr>
<td>2007</td>
<td>261462</td>
<td>626.6</td>
</tr>
<tr>
<td></td>
<td>474</td>
<td>626.6</td>
</tr>
</tbody>
</table>

**Videos - Columbia University Medical Center**

<table>
<thead>
<tr>
<th>Year</th>
<th>Num</th>
<th>Time (Hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>7</td>
<td>51565</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>14.3</td>
</tr>
</tbody>
</table>

**Videos - Computer Science Department**

<table>
<thead>
<tr>
<th>Num</th>
<th>Time (Hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>71.5</td>
</tr>
<tr>
<td>110</td>
<td>93.9</td>
</tr>
<tr>
<td>14</td>
<td>7.9</td>
</tr>
<tr>
<td>62</td>
<td>47.7</td>
</tr>
<tr>
<td>122</td>
<td>94.7</td>
</tr>
<tr>
<td>5</td>
<td>2.9</td>
</tr>
<tr>
<td>493</td>
<td>651.8</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>7</td>
<td>7.8</td>
</tr>
<tr>
<td>899</td>
<td>978.7</td>
</tr>
</tbody>
</table>
History

2005: 10 videos

2006: 100 v.

2007: 1000 v.

UI changes parallel to growth of database
Features over Time

2. VAST MM Orientation

1. Introduction / Background
2. VAST MM Orientation
3. Lecture Videos
4. Presentation Videos
5. Common Indices
6. Video Browsing Characteristic
7. Conclusion

2005 2006 2007 2008
3. Lecture Videos

- **Visual Indices**
  - Create a browser for lecture video content

- **Textual Indices**
  - Create a text index for lectures and courses

- **Task-based Experiments and Evaluation**
  - User study for Visual Indices

- **Course-based Experiments, Evaluation**
  - Why bother with lecture videos?
3.1 Structuring Lecture Videos using Visual Contents (*ICME 2003*)

- Video-taped courses available through “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 keyframes
- Need a more compact, content-directed indexing method for keyframes
Process Overview

1. Video  →  2. "Keyframes"  →  3. Classification by Media Type

5. Visualization and Interface  ←  4. Clustering by Content

3.1 Lecture Videos

- 1. Introduction / Background
- 2. VAST MM Orientation
- 3.1 Visual Indices
- 4. Presentation Videos
- 5. Common Indices
- 6. Video Browsing Characteristic
- 7. Conclusion
Classification of *Media Type*

- Keyframes belong to six Media Types
- Media Type can be easily identified by physical location

<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>
Classification of Media Type

- Examples for each Media Type

- Board
- Podium
- Sheet
- Illustration
- Computer
- Class
Classification of *Media Type*

- Decision Tree (designed by hand)
Clustering by Content Matching

- Snapshot content evolves slowly
- Hand-drawn slides grow monotonically (forward matching)
- Blackboard panels varies dynamically (bi-directional matching)
Clustering by Content Matching

- Content Extraction with background / foreground filters:
  - Color, edges, morphological
- Resulting binary image contains content pixels
Clustering by Content Matching

• Comparisons between sub-windows of keyframes
• Matching by minimizing content pixel differences and sub-window translation vector
Interface and Visualization

Key: Media Types
Topological View
Keyframe Viewer

3.1 Visual Indices
3. Lecture Videos
1. Introduction / Background
2. VAST MM Orientation
3.1 Visual Indices
4. Presentation Videos
5. Common Indices
6. Video Browsing Characteristic
7. Conclusion
Interface and Visualization: Index

• Through analysis: temporal topic model
• Through user study: perception is topologic
• Therefore:
  • Topological View with relative time scale
  • Each topic represented by media icon
  • Interrupted topics reunited by tapered lines
• Verified: users quickly identify key concepts!
Interface and Visualization: Details

• Through user study: need content in context
• Therefore:
  • Keyframe Viewer with full index of key frames
  • Organized by *Media Type* and *Topic*
• Verified: users quickly access full-size images and the video!
Structuring Lecture Videos using Textual Indices (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)
- One lecture: 5k to 14k words \( \approx \) 150k words / course
- Need indices across lectures and courses
- Extract and display structure of entire course using keywords / key 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...

Imperfect Transcript

Interactive Visualization

Lecture Audio

Textbook Index or Manual Index

Index Words and Phrases
Word Pairs
Transcript Generation

- ASR transcript from IBM® ViaVoice®
- Highly compressed lecture video
  - Poor audio quality (although good enough for human perception)
- Transcript:
  - Word Error Rate (WER) of 75% \[ WER = \frac{|insertions| + |deletions| + |replacements|}{|referenceText|} \times 100 \]
  - Stream of words, no punctuation
- Training: little (3%) overall improvement
  - Possibly due to difference in environment: analog vs. digital microphone
  - Number of unique index phrases \( \approx \) same
  - But: difference in identified index phrases 20%
  - Best to combine trained and untrained results
Transcript Generation

• Raw transcript excerpt:

Fees after if Hoffa Allahu of the short and the journalists to leave through today off after next week we have life election Day holiday list possibly visit in the recession possibly for all the week from today in the evening and in France and one later not what what pay half the sugar for slot glasses say and to visit eight half since the Gulf is to with ahead is all Allahu of a of class this risk appears this is a on the real lectures are valid bull on a and talk to see the adapter single constraints for reduced-a any questions rinsed the now is as bad guess they in that strip pipeline is the reading bomb let's go over latency tables again by and did you simulation for floating a kiss with wait-and-see table selects redefine, Aum Costco late as the table when Gillette annual scintillations question as an this got lesser breeds, a to destruction if it paid to complete simulations and should get the DeWitt alert but this action if the look out that tells you what this on in this is what our 8 W dependency as which result in still offer could result best odds of cancer the definition best definition of white and say pop icon and J and I J.R. standard in every is in the chorus in this order = could go unheard of extra were injured in a instructions and under the village general in the work extra clocks cycles between basically the producer in the consumer construction I producer and be a construction K consumer on there are J consumerists result of ought to the very abstract number intervening clocks cycles and this means to visit this it is have fought are these to instructions have to date and the stills still one view of this the is this = first burst in equal number of small cycled suspect in back-to-back did I and J and back-to-back and another in New loop the effect number of in true the name constructions Q bubbling stops as the UN been little formal years which is due in get a at a could sit with the eye and J A rights to register tree leaves from the register to Leslie in Appier back-to-back I list all this still in idea stage and is still cycles that at the Lacy says and is still cycle for extra delay is needed here to make this work that's answered one question second day and I know what to stall at likened politics the problem sell that to
Transcript Generation

• Comparable transcript excerpt:

<table>
<thead>
<tr>
<th>Manual Transcript (129 words)</th>
<th>Automatic Transcript (103 words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>... <strong>deal with</strong> with the data structure like this actually you deal with <strong>it with</strong> with with heaps also so you have some data <strong>structure</strong> right where where items have names and <strong>the question's</strong> how do you how do you get <strong>how do you get</strong> to the items we've actually you you should have asked this <strong>question</strong> already <strong>this</strong> semester right uhm so <strong>and</strong> there this data <strong>structure doesn't provide a way to find something</strong> right like a <strong>BINARY TREE</strong> provides if i'm <strong>looking for 27</strong> in a binary tree you know just <strong>given</strong> the <strong>POINTER to the root of the tree</strong> I have a way to find it right and if you have an array given you know <strong>the name</strong> of the array you have some way to <strong>find</strong> ...</td>
<td>... <strong>deal with</strong> live this <strong>church</strong> is that <strong>CD</strong> in do <strong>it with</strong> wit of the need all sell <strong>Nafta</strong> this <strong>structure</strong> that will write and <strong>assassinations and the question is how you</strong> have to get at it the added that slate on ye shall ask the <strong>question redhead this</strong> vast array of Aum sell <strong>and</strong> it is its <strong>structure doesn't provide a way to find something</strong> like a <strong>BINARY TREE</strong> provides a way of <strong>looking for 27</strong> and by treat it is <strong>given</strong> a <strong>POINTER to the router the treehouse</strong> where it ought and emulate even though <strong>the name Ray Hunt's family finds</strong> ...</td>
</tr>
</tbody>
</table>
Text Filters

• Filter index phrases: structured approach
  • Corpus of expected phrases: 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

• Filter word pairs: unstructured approach
  • Address: speech in lecture fragmented
  • Word pairs in transcript = index words separated by ≤ 10 words

| amortized analysis | (3) (1,2,3) (1,2) |
| account | (1) | call structural |
| method | (2) | call hazard |
| of | (3) | call instruction |
| aggregate analysis of | (1,2) | call compaction |
| call by value | multiple instruction | call step |
| | multiple operation | |
Target Corpus

• Define:
  • *Theme phrases*: General tenor for contents of course, in (> ¼) transcripts
  • *Topic phrase*: Highlight specific topics for lectures, in (< ¼) transcripts
  • *Illustration phrases*: unique terms for examples, hard to identify

• Index Phrases
  • Unique per textbook indices: 253-4701
  • Unique per lecture: 8-98
  • Unique per course: 40-347

Average of 273 transcripts from 11 courses, normalized for number of transcripts.

Average of 6 textbooks, normalized for number of chapters.
Interface Control: Similar to Camera Control

• 3 visualization techniques:
  • Lecture and course text index
  • Lecture to textbook map
  • Lecture similarity map

• 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)
3.2

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 across transcripts
  - Longer blobs for repeating phrases
- Greedy population of space near top

Detail:

- Zoom=1; “highly topic-specific”
- Zoom=13; “course thematic”
Transcripts mapped to textbook chapters
• Rows=transcripts, Columns=chapters
• Match score based on occurrences of terms between transcript and book chapter
• Performance:

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Overall best accuracy over all Zoom levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index Phrases + Word Pairs</td>
<td>0.7</td>
</tr>
<tr>
<td>Word Pairs</td>
<td>0.66</td>
</tr>
<tr>
<td>Word Pairs from G²</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Detail:

<table>
<thead>
<tr>
<th>Course: “Computer Architecture”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
</tr>
<tr>
<td>Incorrect</td>
</tr>
<tr>
<td>Other possibility</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Course: “Analysis of Algorithms”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
</tr>
<tr>
<td>Incorrect</td>
</tr>
<tr>
<td>Other possibility</td>
</tr>
</tbody>
</table>

3.2 Textual Indices

- VAST MM Orientation
- Presentation Videos
- Common Indices
- Video Browsing Characteristic
- Conclusion

- Transcripts mapped to textbook chapters
- Rows=transcripts, Columns=chapters
- Match score based on occurrences of terms between transcript and book chapter
- Performance:
Interface: Chapter Transcript Match

- Chapter-Transcript Matching: Comparison of measures
- Various “zoom” levels, i.e. use of topic vs. thematic terms
- Index Phrases + Word Pairs performs best
  - Marginally better than $G^2$

![Graph showing comparison of measures](image)
Interface: Transcript Similarity

- Cluster similar lectures (transcripts)
  - Dice distance of co-occurrence counts of selected index phrases
    \[ dist(\text{transcript}_i, \text{transcript}_j) = \frac{b + c}{2a + b + c} \]

<table>
<thead>
<tr>
<th></th>
<th>Phrases in ( j )</th>
<th>Phrases not in ( j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrases in ( i )</td>
<td>a</td>
<td>c</td>
</tr>
<tr>
<td>Phrases not in ( i )</td>
<td>b</td>
<td>d</td>
</tr>
</tbody>
</table>

- Multidimensional scaling: \(|\text{lectures}| \rightarrow 2D\)

- Measure user interaction with lecture browser interface
- Participants: 11 students, 1 instructor
- 2 sets of tasks measure effect of familiarity
- **Tasks:**

<table>
<thead>
<tr>
<th>Task</th>
<th>Example formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locate a specific keyframe given a visual example</td>
<td>“Locate this keyframe”</td>
</tr>
<tr>
<td>Locate a keyframe of a media type</td>
<td>“Locate a slide that displays contents from the computer or … shows the professor actively addressing the class”</td>
</tr>
<tr>
<td>Locate a topic (teaching unit)</td>
<td>“Locate a slide from the discussion on …”</td>
</tr>
<tr>
<td>Locate the main topic</td>
<td>“Locate a slide from the main topic for this class”</td>
</tr>
<tr>
<td>Locate the beginning of a topic</td>
<td>“Locate the slide that begins the discussion on …”</td>
</tr>
<tr>
<td>Locating the end of a topic</td>
<td>“Locate the slide that ends the discussion on …”</td>
</tr>
<tr>
<td>Locating the most interaction</td>
<td>“Locate a slide in the portion of the class where the professor switches the most between the different topics on the blackboard”</td>
</tr>
</tbody>
</table>

- Browser used during user study:
  - Keyframe Viewer
  - Keyframe Player
  - Information Mural
  - Topological View
  - External Video Player
  - (User Study Frame)

- From activity logs: intra- and inter-transitions:
  - Information Mural not helpful: remove!
  - Most interaction between Keyframe Viewer and Topological View: combine!

<table>
<thead>
<tr>
<th>Type of Task</th>
<th>Average Time (sec)</th>
<th>Average # Incorrect</th>
<th>Δ Avg. Time (sec)</th>
<th>Δ Avg. # Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locate a keyframe from an example</td>
<td>69.15</td>
<td>0.08</td>
<td>+28.06</td>
<td>-0.17</td>
</tr>
<tr>
<td>Locate a topic (teaching unit)</td>
<td>115.52</td>
<td>0.95</td>
<td>-10.34</td>
<td>+0.24</td>
</tr>
<tr>
<td>Locate the main topic</td>
<td>59.78</td>
<td>0.33</td>
<td>-11.65</td>
<td>0</td>
</tr>
<tr>
<td>Locate the beginning of a topic</td>
<td>89.22</td>
<td>0.92</td>
<td>+11.61</td>
<td>-0.7</td>
</tr>
<tr>
<td>Locating the end of a topic</td>
<td>54.46</td>
<td>1.08</td>
<td>-42.56</td>
<td>-0.83</td>
</tr>
</tbody>
</table>

**Learning Curve:** change between first and second set of questions.

**User Study Frame**

- Keyframe Viewer
- Topological View
- Keyframe Player

• From evaluation: proposed browser
Course-study based Experiments and Evaluation (2007)

- Measure impact of VAST MM as study and resource tool
  - Not a quantitative study of browser performance
  - But: insight into necessity of audio-visual resource tools
- Study: availability during final exam study period (1-2 weeks)
  - 2 courses in Fall 2007:
    - “Computer Architecture”
    - “Programming Languages and Translators”
- 142 students, voluntary participation
  - 91 (64%) did not attempt to use at all
    - Possible reasons: attended lectures
    - Are visual learners (book, notes), etc.
    - Participation not mandatory!
Course-study based Experiments and Evaluation (2007)

- VAST MM tool used features:
  - Text search
  - Browsable cues
  - Annotations
  - Bookmarks
Course-study based Experiments and Evaluation (2007)

• Compare VAST-MM users versus non-VAST-MM users
• Measure difference of midterm ($\alpha$) to final exam score ($\beta$)
  • Normalized difference of standard deviation for each student:
    \[ \Delta_{\text{improvement}} = \frac{\beta_i - \mu_{\text{final}}}{\sigma_{\text{final}}} - \frac{\alpha_i - \mu_{\text{midterm}}}{\sigma_{\text{midterm}}} \]
• Improvement in exam score for VAST-MM users: $\Delta \sigma = 0.29$
• However: difficult to measure isolated effect
• But: comparison between VAST-MM and non-VAST-MM usage for same students under similar conditions
• Therefore: availability in general is favorable
Course-study based Experiments and Evaluation (2007)

• Results:
  • Improvement proportional to usage

![Graph showing Improvement vs. Total amount of time spent with VAST MM (sec)]
4. Presentation Videos

• What are presentation videos?
• Visual Indices
• Audio Indices
• Textual Indices
• Task-based Experiments and Evaluation
• But first: VAST MM Demo

“Oh, have we got a video?”
“YES, WE’VE GOT A VIDEO!!”
- The Young Ones (BBC)
Introduction
(MM 2005, ICME 2007, CIVR 2007)

- Classroom video typically lecture recordings
  - Some editing
  - Semi-professional environment
  - Semi-professional / experienced “actor” (instructor)
- New focus: student presentations and other classroom video material
  - Shift in “actor” from instructor to student
  - Different environment
  - Different use by instructors and students
  - Different challenges for analysis, browsing, and dissemination
- Also: recent lecture videos: more PowerPoint oriented instruction
  - Investigate more general issues to address this
Introduction

• Presentation video:
  • Student(s) present work to class
  • Setting: low-tech classroom
    • No special recording considerations

• Our particular case:
  • Course: *Introduction to Engineering Design*
  • >160 students / semester, 4 course sections
  • Teams of 4-6, ~32 teams / semester
  • Midterm, final presentations, ~16 hours of video
  • Recordings on DV tape, multiple presentations / tape
  • Collection over 5 years: >150 hours of video
Challenges

- Classroom-specific problems:
  - Poor lighting
  - Commotion due to lack of “stage”
  - Varying audio quality
  - Amateur camera operator (intentional)
  - No editing (too expensive, no merit)
Visual Segmentation

**Objectives**
- Determine scenes of visual similarity
- Condense visual information for browser

**Problem**
- Unedited video, thus no scene cuts

**Approach**
- Detect two types of predominant visual change events
Visual Segmentation

1. Abrupt change:
   • E.g., electronic slide change
   • Pixel intensity difference between consecutive frames

2. Gradual change:
   • E.g., person walking in/out, camera pan/zoom
   • Histogram change between distant frames (4 secs)
   • Combination of the two:
     • Measure of degree of change (not binary)
       →Customizable parameter in UI
Visual Feature Control: Scenes

- Adjustable parameter: visual granularity
  - Fine granularity: slight distinct visual changes, more keyframes
  - Coarse granularity: more distinct visual changes, fewer keyframes
Visual Feature Control: Zoom

- Adjustable parameter: temporal zoom
  - Change level of detail for video summary
  - Decrease/increase amount of displayed information
Speaker Segmentation

- **Objectives**
  - Determine audio scene changes
  - Provide visual index of speakers

- **Approach**
  - Speaker Segmentation
    - Use well-accepted Bayesian Information Criterion with MFCC features; high accuracy
4.3 Audio Indices

Speaker Clustering (ICME 2008)

- Objectives
  - Provide speaker index
  - Visually emphasize speaker recurrences and dialogs
  - Shown here: Proposed graphs

- Approach: comparison of speakers via symmetric KL

Without visual speaker labeling

With visual speaker labeling
Symmetric KL (KL2)

• 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
Simulation of KL2

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

- 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
Visual Face Index

- Speaker Index
  - In lieu of face recognition
  - Manual extraction in this iteration for experiments
    - Are speaker indices useful?
  - User study on various types of face representations (results shown later)
Text Augmentation

**Objective**
- Provide searchable index
- Visually emphasize keywords and key phrases

**Challenge**
- No speaker model training (> 160 students per semester)
- No language model training (highly varying styles among > 160 students)
## Text Filter

### Approach

- Filter “good” content using external corpus (presentation slides)
  - Determine descriptive phrases in slides with WordNet
- Rank filtered words/phrases
  - Nouns more descriptive, thus rank higher
  - Longer phrases more descriptive, thus rank higher
  - No stop words
  - Phrases and rank visualized in UI

<table>
<thead>
<tr>
<th>4. Presentation Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction / Background</td>
</tr>
<tr>
<td>2. VAST MM Orientation</td>
</tr>
<tr>
<td>3. Lecture Videos</td>
</tr>
<tr>
<td>4.4 Textual Indices</td>
</tr>
<tr>
<td>5. Common Indices</td>
</tr>
<tr>
<td>6. Video Browsing Characteristic</td>
</tr>
<tr>
<td>7. Conclusion</td>
</tr>
</tbody>
</table>
4.4 Visual Feature Control: Context

- Adjustable parameter: Text Context
  - Group temporally close words/phrases (x seconds)
  - Visually isolates themes

- Measure quantitatively through targeted tasks
- Extensive: 3 years of user studies
- Improvement of VAST MM based on user study results
- Tasks:

<table>
<thead>
<tr>
<th>Task</th>
<th>Example formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locate familiar content</td>
<td>“Find your appearance during your team’s presentation”</td>
</tr>
<tr>
<td>Locate familiar content</td>
<td>“Find the beginning of your team’s presentation”</td>
</tr>
<tr>
<td>Locate familiar content</td>
<td>“Find your team’s discussion on functional requirements”</td>
</tr>
<tr>
<td>Locate unfamiliar content</td>
<td>“Find the presentation related to building a greenhouse for a local school”</td>
</tr>
<tr>
<td>Summarize unfamiliar content</td>
<td>“Locate the presentation in video X between time Y-Z. Summarize the presentation using only the displayed keywords and phrases”</td>
</tr>
</tbody>
</table>

• Measures:
  • Completion rate: successful completion of task, i.e. correct answer
  • Time: duration of task
  • Accuracy:
    • For search tasks: distance between expected and recorded answer
    • For summary task: subjective measure of semantic similarity

• Quantitative results:
  • Significant improvements in difficult search tasks

• Qualitative results (anecdotal)
  • Instructional staff notes: Material available to students helps prepare them better for their own “first” presentation
4.5 Task-based Evaluation

Evaluation: Structure In Video

- Measure impact of structural versus non-structural cues

Structural:
- Presentation segmentation (via hand clap detector)
  - chapters of book
- Keyframes
  - sections in chapter of book

Non-structural:
- Streaming video
- Text

<table>
<thead>
<tr>
<th></th>
<th>Keyframe</th>
<th>Streaming</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presentation Seg.</td>
<td>Structured</td>
<td>Semi-structured</td>
</tr>
<tr>
<td>Text</td>
<td>Semi-structured</td>
<td>Unstructured</td>
</tr>
</tbody>
</table>

- 4 groups of participants:
- Best performance: structural
- Therefore: structure from video important!
# Evaluation:

**Structure In Video**

## Participant sample size:

<table>
<thead>
<tr>
<th></th>
<th>Keyframe</th>
<th>Streaming</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presentation Seg.</td>
<td>40</td>
<td>39</td>
</tr>
<tr>
<td>Text</td>
<td>37</td>
<td>40</td>
</tr>
</tbody>
</table>

## Task: “Find the beginning of your first (or only) appearance in which you spoke during the presentation”

<table>
<thead>
<tr>
<th></th>
<th>Keyframe</th>
<th>Streaming</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presentation Seg.</td>
<td>31.78</td>
<td>72.43</td>
</tr>
<tr>
<td>Text</td>
<td>38.94</td>
<td>87.80</td>
</tr>
</tbody>
</table>

## Task: “Find the beginning on your team’s discussion on Functional Requirements”

<table>
<thead>
<tr>
<th></th>
<th>Keyframe</th>
<th>Streaming</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presentation Seg.</td>
<td>57.38</td>
<td>125.77</td>
</tr>
<tr>
<td>Text</td>
<td>68.14</td>
<td>106.58</td>
</tr>
</tbody>
</table>

## Task: “Find the presentation on <topic XYZ> in any of the provided videos”

<table>
<thead>
<tr>
<th></th>
<th>Keyframe</th>
<th>Streaming</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presentation Seg.</td>
<td>190.24</td>
<td>348.70</td>
</tr>
<tr>
<td>Text</td>
<td>217.69</td>
<td>353.22</td>
</tr>
</tbody>
</table>

## Task: “Locate presentation <X> and summarize using text index”

<table>
<thead>
<tr>
<th></th>
<th>Keyframe</th>
<th>Streaming</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presentation Seg.</td>
<td>59.95</td>
<td>126.44</td>
</tr>
<tr>
<td>Text</td>
<td>48.59</td>
<td>71.30</td>
</tr>
</tbody>
</table>
Evaluation: Streaming Video

• Availability of Streaming Video is counter-productive

• 2 groups of participants:
  1) Only index cues
  2) Index cues + video

• Results:
  • Both groups: similar completion rate
  • Without streaming video: requires significantly less time
  • With streaming video: students pay less attention to index cues, fall back on familiar medium

• Therefore:
  • Index cues better for search and retrieval!
  • Disable streaming video during search?
Evaluation: Streaming Video

• Results:

<table>
<thead>
<tr>
<th></th>
<th>Video Streaming</th>
<th>No Video Streaming</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Completion</td>
<td>Duration (sec)</td>
</tr>
<tr>
<td>Spring 2005</td>
<td>90%</td>
<td>84.81</td>
</tr>
<tr>
<td>Fall 2005</td>
<td>89%</td>
<td>126.78</td>
</tr>
<tr>
<td>Spring 2006</td>
<td>89%</td>
<td>131.71</td>
</tr>
</tbody>
</table>
Evaluation: Face Indices

- Tasks pertaining to locating people benefit from face index
- Test 4 configurations

- Best performance: Lg. head + profile = most visual information

![Effect of various Face Indices](image)

4.5 Task-based Evaluation

4. Presentation Videos

1. Introduction / Background
2. VAST MM Orientation
3. Lecture Videos
4.5 Task-based Evaluation
5. Common Indices
6. Video Browsing Characteristic
7. Conclusion

<table>
<thead>
<tr>
<th>Task Duration (sec)</th>
<th>Small face</th>
<th>Two Small Faces</th>
<th>Large head/shoulder</th>
<th>Two Large Faces</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find self appearance</td>
<td>140</td>
<td>120</td>
<td>160</td>
<td>130</td>
</tr>
<tr>
<td>Find other person's appearance</td>
<td>130</td>
<td>110</td>
<td>150</td>
<td>120</td>
</tr>
<tr>
<td>Average</td>
<td>135</td>
<td>125</td>
<td>155</td>
<td>125</td>
</tr>
</tbody>
</table>
Evaluation: VAST MM Improvements

- Improvement of VAST MM over time
- Best measure: difficult search task (finding unfamiliar content)
  - Completion rate increases: 57% to 97%
  - Task duration decreases: 436 sec to 128 sec

![Graph showing successful search task completion rate over time](image1)

Completion rate increases

![Graph showing search task duration over time](image2)

Task duration decreases
5. Common Indices

- Temporal speech-text alignment
- Keyword / key phrase ranking and UI
- Video annotations
- Taxonomy browser
5.1 Temporal Speech-Text Alignment

(ICME 2007)

• Speech-text alignment
  • ASR transcripts: stream of text, no timestamps
  • Need timestamps to link ASR text to correct location in video browser
  • Should be extractable from ASR tools, but it isn’t!
    • Not in IBM ViaVoice
    • Not in Dragon NaturallySpeaking

• Therefore: post process text-speech alignment
5.1 Speech-Text Alignment

1. ASR Silence Filter
2. Noise Filter
3. Align phonemes

<table>
<thead>
<tr>
<th>Phonemes:</th>
<th>Audio Phoneme - Text Phoneme Alignment:</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA AH AA AH AE AS AH S SH S AE S IH SH AH</td>
<td></td>
</tr>
<tr>
<td>1.03 1.06 1.27 1.33 1.39 1.97 1.97 2 2.16 2.2 2.23 2.3 2.4 2.47 2.49 2.77 2.9</td>
<td></td>
</tr>
</tbody>
</table>

Transcript Text:

- Transcript: "align this transcription"
- Words to Phonemes:
  1. AH0 L AY1 N
  2. DH IH1 S
  3. T R AE2 N S K R ..
  4. IH1 P SH AH0 N
- Phoneme Selection:
  1. AH AH
  2. IH S
  3. AE S IH SH AH

Monophthongs: IY (beet), IH (bit), EH (bet), AE (bat), AH (ab-ove), UW (boot), UH (book), AA (father), ER (bird), AO (bought)
Fricatives: SH (assure), S (sign)

Diphthongs: AW (out) → AH, AY (fîve) → AH, EY (day) → AE, OW (crow) → UH, OY (boy) → AO

Fricatives: Z (resign) → S
Affricates: CH (church) → SH
## Results

- **Reasonable accuracy:**

<table>
<thead>
<tr>
<th>Video; Transcription generation</th>
<th>Features</th>
<th>Length</th>
<th>Avg. Matching Error (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture; manual</td>
<td>Single speaker, one long break (504 sec)</td>
<td>1:48:21</td>
<td>3.9</td>
</tr>
<tr>
<td>Lecture; automatic</td>
<td>Single speaker, one long break (504 sec)</td>
<td>1:48:21</td>
<td>7.7</td>
</tr>
<tr>
<td>Student Presentation; automatic</td>
<td>31 speakers, 6 Q&amp;A sessions of varying durations (30 – 300 sec)</td>
<td>1:15:12</td>
<td>6.43</td>
</tr>
<tr>
<td>Student Presentation; automatic</td>
<td>10 speakers, 2 Q&amp;A session of durations 60 sec and 185 sec</td>
<td>0:22:32</td>
<td>26.73</td>
</tr>
</tbody>
</table>
5.2 Keyword / Key phrase Ranking and UI (SIGIR 2007)

• Rank words/phrases by uniqueness and semantic saliency
• Emphasize rank in UI
• Combine various WordNet measures for a phrase:
  • Number of terms in a phrase: more saliency if more terms
    • “gulf” vs. “gulf of Mexico”
  • Synonymous senses: more salient if fewer
  • Distance to root (WordNet hierarchy): more salient if more distant
  • Nouns are more descriptive than verbs: rank higher
User Interface

- UIs with and without rank emphasized
- Limited user study on perception
  - UI with rank better highlights key concepts
  - Summarization task completed faster and with higher accuracy
Annotations / Bookmarks

- Make videos more interaction
- Public annotations: disseminate feedback
- Private bookmarks: mark specific content, not the entire video
5.4 Taxonomy Browser

- With growing video collection:
  - Need organization (categorical)
  - Search is ok to find specific content; does not address “browsing”

- Therefore: taxonomy browser
  - Temporal dimension: categories over time
    - E.g. “Course W4771: Fall ‘03, Fall ‘04, …”
  - Spatial dimension: categories have similarities
    - E.g. “Final Presentation ~ Midterm Presentation !~ Client Meeting”

- Taxonomy arbitrary, we build it manually
3D Cylinder UI

• We model categories in 3D cylinder
  • Time: length of cylinder, divided into sections
    • Sections arbitrary; we chose years; semester granularity perhaps better!
  • Categories within in a section (year): labels on cylinder surface
    • Labels are spread on surface according to similarity
    • Distance metric: manually determined
  • Categories across sections:
    • Linked visually by curve on cylinder surface
    • Spread out according to similarity
    • Distance minimization: iterative

• User Interaction
  • Mouse – Press and Drag
  • Vertical motion: rotates cylinder
  • Horizontal motion: pans cylinder
3D Cylinder UI

5.4 Taxonomy Browser

Student Presentation Midterm
Student Presentation Final

Mouse - Press and Drag

Perspective: Most Depth Least
Number of Faces: Least Smoothness Most
Evaluation

• Very preliminary (in 1 semester, not main focus of user study)

• Quantitative:
  • Task: “Select / navigate to category XYZ”
    • E.g.: “Select the category for the earliest final presentation videos in E1102”
  • Successive completion times:
    • 56.5, 38.8, 33.5, 27.6, 20.6 seconds
  • Constant decrease in time suggests: need more tasks

• Qualitative:
  • Bi-modal: either loved it (“fun”, “cool”) or hated it (“ugly”, “too difficult”)
  • Ironically: course (for user study) focused strongly on 3D modeling
6. Video Browsing Characteristics

- Why not use a standard video player?
- Browsing styles
- Comparison to VAST MM
Introduction
(submitted: MM 2008)

• Qualify/quantify user interaction with standard video players
• Compare VAST MM to standard video player
  • Through difficult search task: finding unfamiliar content
• User study
  • Large database: 170 hours of video data
  • 2 groups of participants:
    • (1) use standard video player
    • (2) use VAST MM
  • 9 search tasks, random order for each participant
  • Answer as many in 30 minutes as possible
User Interfaces for Study

- Video players / browser used to study behavior
  - Simulate standard video player: play, pause, stop, location slider, keyframe browsing via location slider
User Study Tasks

- Tasks of varying difficulty
  - Common versus rare contents
  - Entire video database = 611,388 seconds

- Tasks:

<table>
<thead>
<tr>
<th>Type of Search</th>
<th>Search task “Find video content for presentations of or relating to …”</th>
<th>Time (sec) of Video Content</th>
<th>Percentage of Library Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific Presentation</td>
<td>Musical Device for People with Cerebral Palsy</td>
<td>1,860</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>Coogan's Restaurant - Food Waste to Energy</td>
<td>1,860</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>CU Study Away - A Website for the Study Abroad Program</td>
<td>2,160</td>
<td>0.4</td>
</tr>
<tr>
<td>Specific Client</td>
<td>The MTA (Metropolitan Transportation Authority)</td>
<td>3,600</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>The 125th Street BID</td>
<td>7,200</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>Dr. Gil Lopez of CMSP Math Scales</td>
<td>14,400</td>
<td>2.4</td>
</tr>
<tr>
<td>Project Category</td>
<td>Information Technology (Web sites, Database projects, Handheld devices, etc.)</td>
<td>106,200</td>
<td>17.4</td>
</tr>
<tr>
<td></td>
<td>Architectural Design (e.g. Lab space design, Office design, etc.)</td>
<td>117,000</td>
<td>19.1</td>
</tr>
<tr>
<td></td>
<td>Disabilities</td>
<td>223,200</td>
<td>36.5</td>
</tr>
</tbody>
</table>
Results: Browsing Styles

- Standard video player: 2 browsing styles
  - Visual: users skim keyframes
  - Audio-visual: users view short clips of video

- Observations:
  - Distribution of users almost equal
  - Users strictly adhere to one method
  - Audio-visual browsing: better performance

<table>
<thead>
<tr>
<th></th>
<th>Users</th>
<th>Attempted Tasks</th>
<th>Task Time</th>
<th>Correct Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio-Visual Browsing</td>
<td>39</td>
<td>97</td>
<td>531 sec</td>
<td>36</td>
</tr>
<tr>
<td>Visual Browsing</td>
<td>40</td>
<td>56</td>
<td>846 sec</td>
<td>15</td>
</tr>
</tbody>
</table>
Results: Browsing Styles

- Visual-only vs. Audio-visual browsing:
  - Similar behavior at different temporal scales
  - Visual-only = RED, Audio-Visual = BLUE
6. Index-cue Browser versus Standard Video Player

• In theory:
  • Without indices
    • Search = random selection and verification
    • Hit rate = inversely proportional to amount of matching content
  • With indices
    • Hit rate constant, regardless of amount of matching content

• Verification through user studies:

![Task Time: Index-cue driven browser vs. video player](image)

- Index-cue driven browser
- Standard video player
- Fit to power law
- Fit to constant
7. Conclusion / Future Work

- Much to be done …
- Many people to convince …
Conclusion

• Demonstrated approaches for video indexing:
  • Traditional lecture videos (visual, textual)
  • Presentation videos (visual, audio, textual)

• Introduced novel browser
  • Provides search and browsing features
  • Ties together multi-modal cues

• Through user studies:
  • Tools improve and help in content retrieval of large libraries

• Through feedback (students and instructors):
  • Availability enhances student learning (self- and peer-evaluation)
  • Good resource for study
• Tested on third genre: team meetings, team interaction
  • Challenges:
    • Very candidly captured; qualities vary substantially
    • Unlikely index: Text; text filters not well-defined
    • Likely index: Faces; although needs automation
Future Work

• Further automate successful indices
  • In particular: face extraction and display

• New indices
  • Find speaker based on example data (e.g. speaker clustering)
  • Phoneme-based text search

• Mapping between videos and external corpora
  • Map structure (e.g. textbook to lecture)
  • Map information (can a video’s content suggest external material?)

• Cross-video threads
  • Recurring contents (can a video’s content suggest another video?)
  • Recurring actors

• Apply to other video genres
Thank you!