Augmented Segmentation and Visualization for Presentation Videos

Alexander Haubold and John R. Kender
Department of Computer Science
Columbia University, New York
{ahaubold,jrk}@cs.columbia.edu
Overview

• Motivation
• Characteristics of Presentation Video
• Segmentation (Audio, Visual)
• Segmentation (Combined Audio-Visual)
• Text Augmentation
• Interface
• Demo
• User Study
• Conclusion
• Future Investigations
Overview

- Motivation
- Characteristics of Presentation Video
- Segmentation (Audio, Visual)
- Segmentation (Combined Audio-Visual)
- Text Augmentation
- Interface
- Demo
- User Study
- Conclusion
- Future Investigations
Motivation

- Videos of student team presentations
- 1 semester ≈ 160 students, 30 teams, 8 hours of video for midterm presentations
- How to best review?
- Need automatic index for videos
- Need visual browser for searching
Overview

- Motivation
- Characteristics of Presentation Video
  - Segmentation (Audio, Visual)
  - Segmentation (Combined Audio-Visual)
  - Text Augmentation
  - Interface
  - Demo
  - User Study
  - Conclusion
  - Future Investigations
Characteristics

- Multiple speakers: $\approx 5 / \text{team, } \approx 20 / \text{hour}$
- 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
Characteristics

- **Thematic Phrases**
- **Topic Phrases**
- **Video Segment**
- **Audio ASR**
- **Align audio/ASR**
- **Thumb**
- **Database**

**Pres. Video**

**Video**

**Audio**

**ASR**

**Segment**

**Theme Phrases**

**Topic Phrases**

**Database**

**UI**
Overview

• Motivation
• Characteristics of Presentation Video
  • Segmentation (Audio, Visual)
  • Segmentation (Combined Audio-Visual)
• Text Augmentation
• Interface
• Demo
• User Study
• Conclusion
• Future Investigations
Segmentation (Audio)

- Identify audio segments for each student
- MFCCs for representing features of speech
- Bayesian Information Criterion detects speaker changes
- Results encouraging, even for varying audio quality

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th># Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>88.5%</td>
<td>95.7%</td>
<td>395</td>
</tr>
</tbody>
</table>
**Segmentation (Visual)**

- **Boundaries from non-overlapping sources:**
  - Presentation slide changes
    - Not all presentations have slides
  - Speaker gesture changes
    - Long-term change in speaker pose
    - Reconfiguration of speaker position
    - Amount of gesture

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th># Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>89.4%</td>
<td>82.7%</td>
<td>594</td>
</tr>
</tbody>
</table>
Overview

• Motivation
• Characteristics of Presentation Video
• Segmentation (Audio, Visual)
  • Segmentation (Combined Audio-Visual)
• Text Augmentation
• Interface
• Demo
• User Study
• Conclusion
• Future Investigations
Segmentation (Both)

• 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
Segmentation (Both)

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th># Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>89.3%</td>
<td>92.7%</td>
<td>710</td>
</tr>
</tbody>
</table>

- Compare to separate segmentations w.r.t. presentation units:

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>51.3%</td>
<td>53.2%</td>
</tr>
<tr>
<td>Video</td>
<td>66.6%</td>
<td>69.2%</td>
</tr>
</tbody>
</table>
Overview

- Motivation
- Characteristics of Presentation Video
- Segmentation (Audio, Visual)
- Segmentation (Combined Audio-Visual)
- Text Augmentation
  - Interface
  - Demo
  - User Study
  - Conclusion
  - Future Investigations
Text Augmentation

• ASR transcript from IBM® ViaVoice®
  – Poor audio quality
  – No training (would require 160 / semester)
  – Word Error Rate of 75%

• Apply 2 filters
  – Manually assembled list of “theme phrases”
    • Phrases / titles of required sections
  – Automatic list of “topic phrases” from presentation slides (if available)
    • Appear in presentation AND transcript
Text Augmentation

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>

Task Objective Demo Team development Gantt chart Statement Limitations Deliverables Team process Future directions Solutions Implementation Continuity Tasks performed Functional Requirements Schedule Goal Constraints Project goals Design Constraints Requirements Future Chart Problem statement Continuity Plan Prototype

Topic Phrases

family data like wheelchairs more issues david third people well working leave design use design

functional functional
goal solutions

make live home working like like like hope home family

use intent cerebral policy use make sure design

house simple live kids working access home

movements working simple more family working

house home details hope
Overview

• Motivation
• Characteristics of Presentation Video
• Segmentation (Audio, Visual)
• Segmentation (Combined Audio-Visual)
• Text Augmentation
• Interface
• Demo
• User Study
• Conclusion
• Future Investigations
Interface

- List of Videos
- Zoomable Summary
- Video Playback

- Thumbnails
- Timeline
- Audio, video tracks
- Text tracks
Interface: Timeline

- 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
Overview

• Motivation
• Characteristics of Presentation Video
• Segmentation (Audio, Visual)
• Segmentation (Combined Audio-Visual)
• Text Augmentation
• Interface
• Demo
• User Study
• Conclusion
• Future Investigations
Overview

- Motivation
- Characteristics of Presentation Video
- Segmentation (Audio, Visual)
- Segmentation (Combined Audio-Visual)
- Text Augmentation
- Interface
- Demo

- User Study
- Conclusion
- Future Investigations
User Study

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

### Task List

<table>
<thead>
<tr>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find your appearance during presentation</td>
</tr>
<tr>
<td>Find beginning of your team’s presentation</td>
</tr>
<tr>
<td>Find you 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)
Overview

- Motivation
- Characteristics of Presentation Video
- Segmentation (Audio, Visual)
- Segmentation (Combined Audio-Visual)
- Text Augmentation
- Interface
- Demo
- User Study
- Conclusion
- Future Investigations
Conclusions

• 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
Overview

• Motivation
• Characteristics of Presentation Video
• Segmentation (Audio, Visual)
• Segmentation (Combined Audio-Visual)
• Text Augmentation
• Interface
• Demo
• User Study
• Conclusion

• Future Investigations
Future Investigations

• Active displays
  – What you see on UI must be clickable
• Topological grouping
  – Temporally group similar audio/visual sources
• Speaker gesture
  – Classification and labeling of speakers
• Annotation tool
  – Instructors / students annotate presentations
Thank you!

Questions / Answers?