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 $\approx$ 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

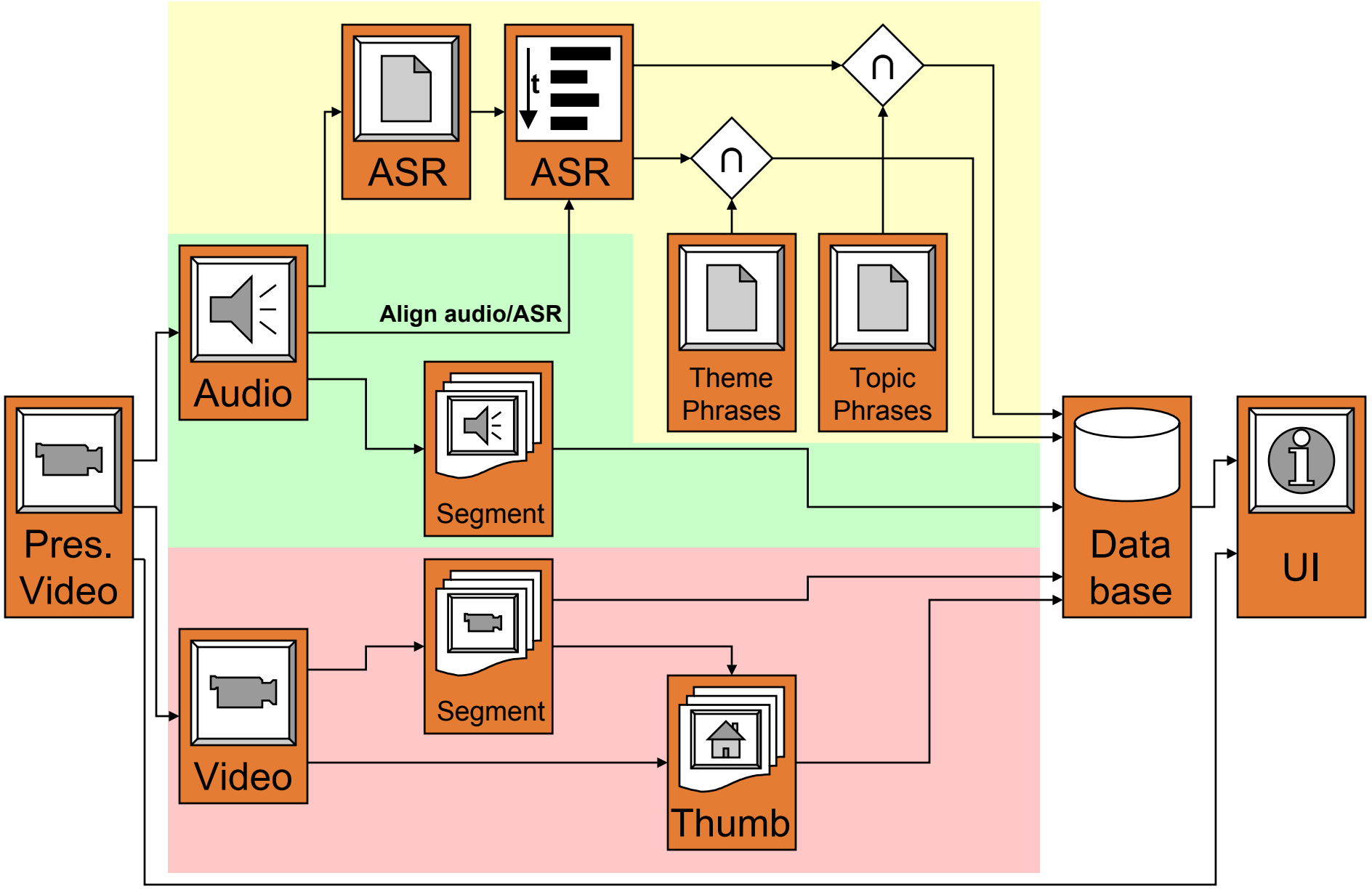
Overview

- 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 / team, $\approx$ 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

Characteristics



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

Precision	Recall	# Segments
88.5%	95.7%	395

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

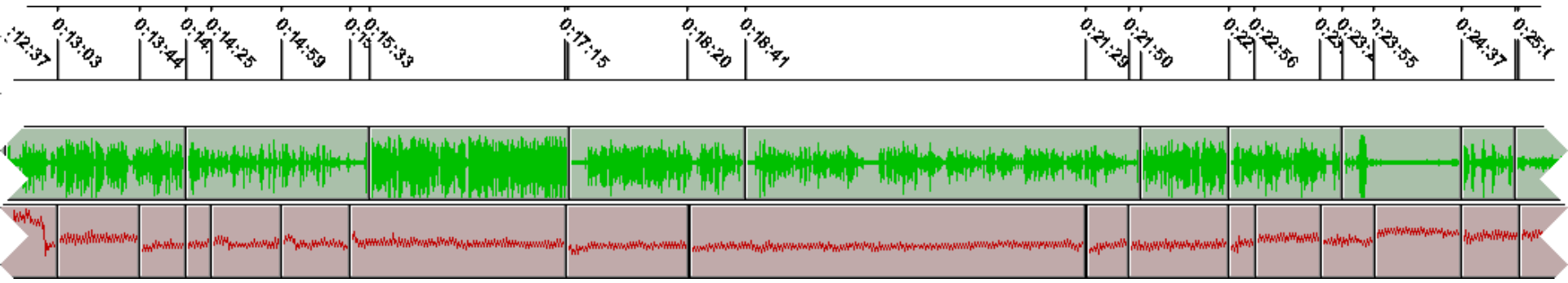
Precision	Recall	# Segments
89.4%	82.7%	594

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)

Precision	Recall	# Segments
89.3%	92.7%	710

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

	Precision	Recall
Audio	51.3%	53.2%
Video	66.6%	69.2%

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

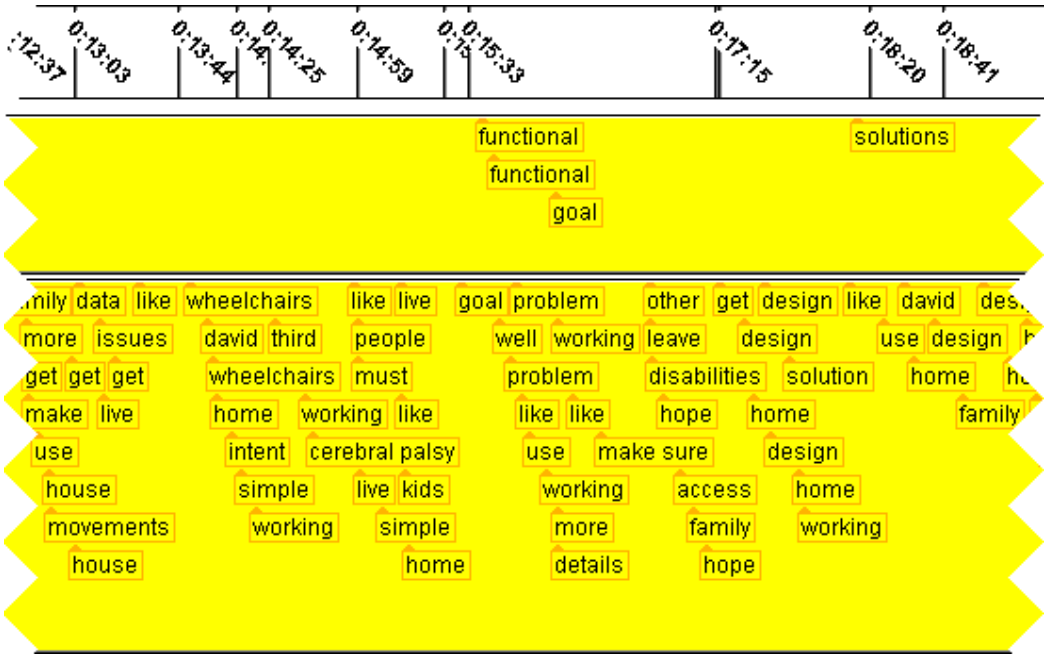
Text Augmentation

Theme Phrases:

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

Theme Phrases →

Topic Phrases →



Overview

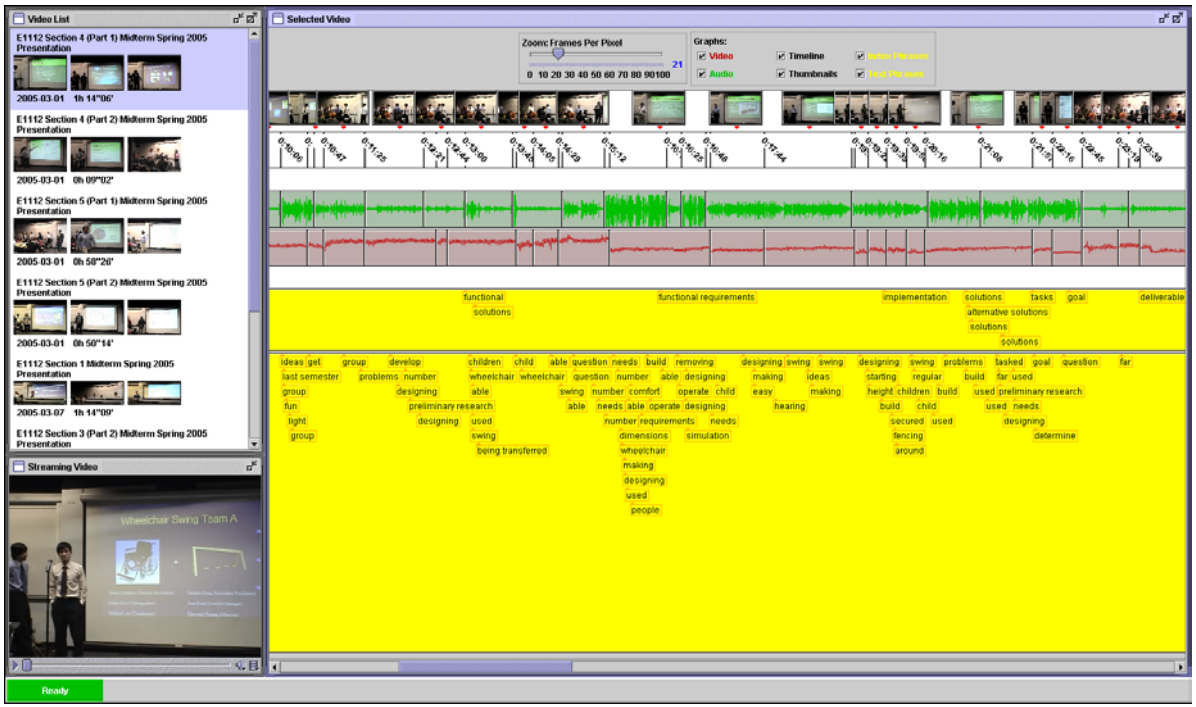
- Motivation
- Characteristics of Presentation Video
- Segmentation (Audio, Visual)
- Segmentation (Combined Audio-Visual)
- Text Augmentation
- Interface
- Demo
- User Study
- Conclusion
- Future Investigations

Interface

- List of Videos
- Zoomable Summary
- Video Playback

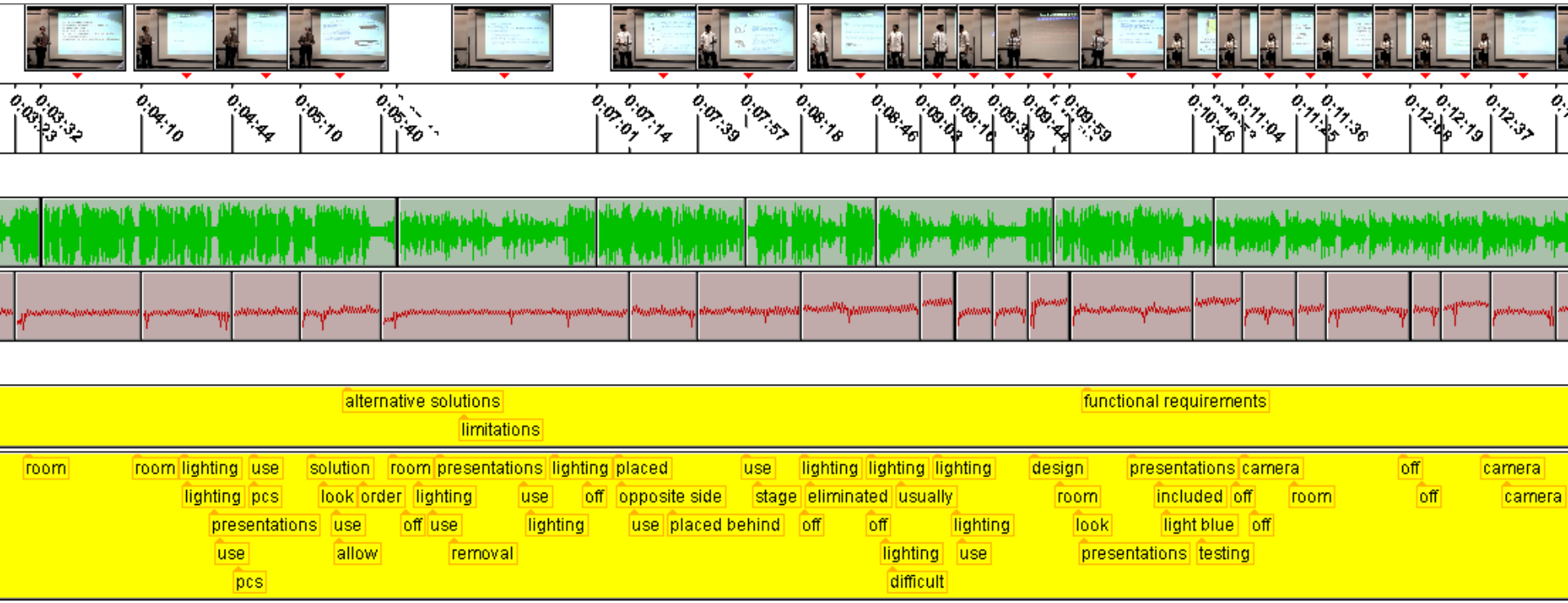
→

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



Interface: Timeline

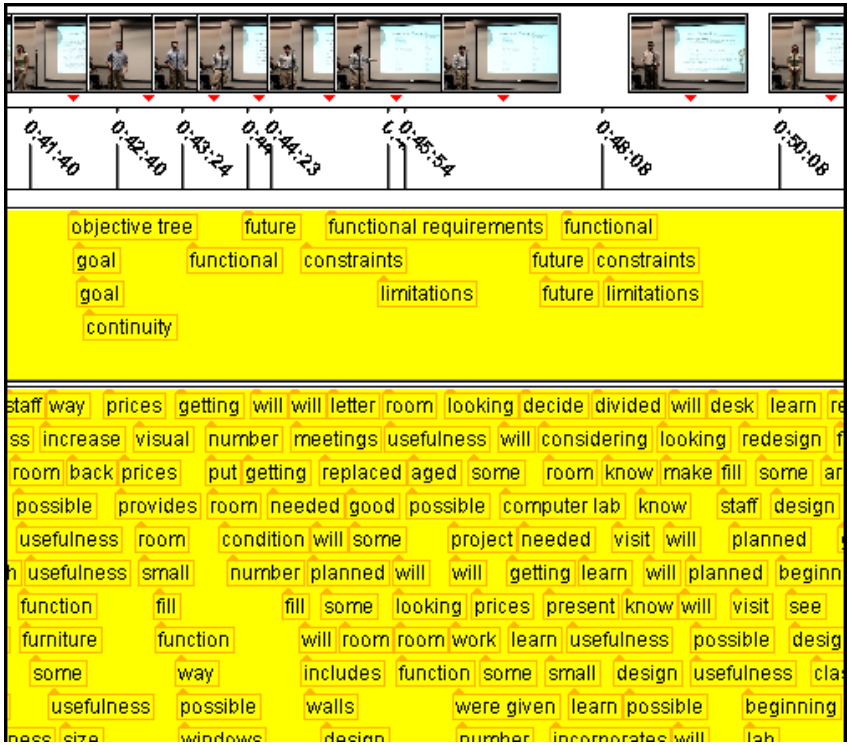
- Portrait notebook-style not well received
- Re-modeled to horizontal continuous timeline



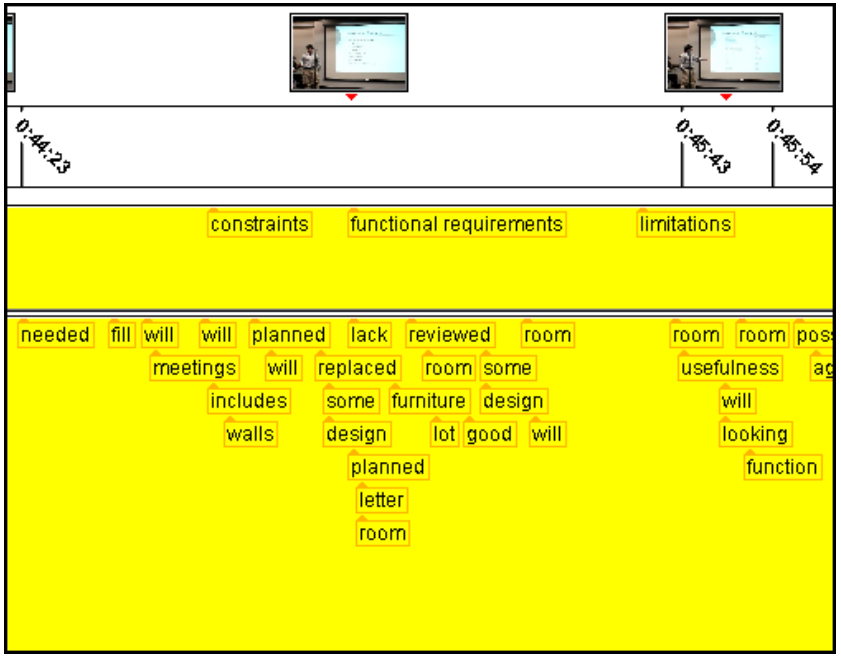
Interface: Text Graph

- 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

Find your appearance during presentation
Find beginning of your team's presentation
Find you team's discussion on topic X
Find presentation Y (Y of different team & class)
Summarize segment using only text

- 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?