VAST MM: Multimedia Browser for Presentation Video

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

Presented by: Apostol (Paul) Natsev
IBM T.J. Watson Research Center
Overview

• Introduction
  – Motivation
  – Background
  – Challenges
  – Objectives

• Indexing
  – Visual Segmentation
  – Speaker Segmentation
  – Text Index

• User Interface

• Evaluation

• Conclusion / Future Work
Overview

- Introduction
  - Motivation
  - Background
  - Challenges
  - Objectives

- Indexing
  - Visual Segmentation
  - Speaker Segmentation
  - Text Index

- User Interface

- Evaluation

- Conclusion / Future Work
Introduction

• 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
Introduction: Motivation

- Classroom video beyond lectures are important resource
- Presentation video focus on students
  - Self evaluation: presentation style, fluency, etc.
  - Team evaluation: team dynamics
  - Peer evaluation: compare to / learn from others
  - Important for engineering education
    - ABET (Accreditation Board for Engineering and Technology)
    - Instruction on professional skills, e.g. presentation, communication
- Archive
  - Long-term progress (student- and course-specific)
Introduction: Background

• 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
Introduction: 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)

In-class presentation with Q&A
Introduction: Objectives

• Video browser for classroom material
  – Easy accessibility (WWW)
  – Integrated (video, visual browsing cues, search)
  – Interactive UI elements
  → Java app. with platform-independent video player

• Automatic, inexpensive indexing of content
  – Manual indexing burdensome
  – Material does not merit manual effort
  – But: long videos require content browsing cues

• Some search functionality
  – Reasonable cues from speech
Overview

• Introduction
  – Motivation
  – Background
  – Challenges
  – Objectives

• Indexing
  – Visual Segmentation
  – Speaker Segmentation
  – Text Index

• User Interface
• Evaluation
• Conclusion / Future Work
Indexing: 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
Indexing: 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
Indexing: 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
  – Visual Speaker Index
    • Extract face regions from video segments
Indexing: Speaker Segmentation

- 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
Indexing: Text Index

- **Objective**
  - Provide keywords / phrases as content index
  - Use ASR (IBM ViaVoice) for automatic transcription

- **Problem**
  - Many speakers (> 160 per semester)
  - Speaker models not available / too burdensome
Indexing: Text Index

• **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
Overview

• Introduction
  – Motivation
  – Background
  – Challenges
  – Objectives

• Indexing
  – Visual Segmentation
  – Speaker Segmentation
  – Text Index

• User Interface

• Evaluation

• Conclusion / Future Work
User Interface

• Video Content Browser
  – Video scenes (dynamic visual change) indicated by thumbnails
  – Speaker scenes indicated by faces
  – Keywords/phrases ranked by descriptiveness and frequency of use
  – Streaming video (MPEG1)
    • Pure Java implementation (server/client) instead of JMF (JMF unreliable, platform-dependent, not readily available)
    • Cross-platform: Windows, MacOS, Linux, Solaris
  – Text Search over filtered text and raw transcripts
User Interface

• Adjustable parameter: Visual Granularity
  – Coarse granularity: more distinct visual changes
  – Decrease/increase number of thumbnails
User Interface

- Adjustable parameter: Zoom
  - Change level of detail for video summary
  - Decrease/increase amount of displayed information
User Interface

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

• Introduction
  – Motivation
  – Background
  – Challenges
  – Objectives

• Indexing
  – Visual Segmentation
  – Speaker Segmentation
  – Text Index

• User Interface

• Evaluation

• Conclusion / Future Work
Evaluation

• Generally not straightforward to measure
• Some quantitative results
  – End-of-semester user studies
  – Targeted tasks
  – 4 semesters (2 years), 598 students
  – Usage logs
• Some qualitative results
  – Surveys completed by students
  – Anecdotal responses from instructors
Evaluation

- Quantitative Setup
  - Students complete targeted tasks
    - Search for familiar content
      e.g., “Find your first appearance”
    - Search for unfamiliar content
      e.g., “Find presentation on topic XYZ”
    - Summarize presentations using keywords/phrases
  - 8 hours of video from on-going semester
    - 2 hours are familiar content (student’s course section)
    - 6 hours are unfamiliar (3 parallel course sections)
  - Measures
    - Completion rate: completed vs. skipped (~ frustration rate)
    - Duration of completion
    - Accuracy: distance of response to groundtruth (in seconds)
Evaluation

• Some quantitative results over 3 semesters (Continual improvement of indexing and user interface)
  – Reasonable and promising improvement

<table>
<thead>
<tr>
<th>All Tasks</th>
<th>Fall 2005</th>
<th>Spring 2006</th>
<th>Fall 2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completion</td>
<td>81%</td>
<td>83%</td>
<td>92%</td>
</tr>
<tr>
<td>Average Duration</td>
<td>120.11 sec</td>
<td>126.66 sec</td>
<td>100.22 sec</td>
</tr>
</tbody>
</table>
Evaluation

• Finding unfamiliar content
  “Find presentation on subject XYZ”
  – Most difficult task, over 8 hours of video
  – Few hints of answer (if any) in visual cues
  – Interesting effect when text search available (*)
    • Text search isolates likely video(s)
    • If unable to locate answer within, blindly trust search result
      → Student selects correct video, but answer within is far off
      → Increase in rate of completion, decrease in avg. accuracy

<table>
<thead>
<tr>
<th>Finding unfamiliar content in 8 hours of video (most difficult task)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measure</strong></td>
</tr>
<tr>
<td>Completion</td>
</tr>
<tr>
<td>Average Duration</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
</tbody>
</table>
Evaluation

• Visual speaker index
  – Determine representation of faces for browsing
  – User task: find student given example of face
    – 4 representations (randomly assigned in user study)
      1. Head/Shoulder + Profile/Shoulder
      2. Head/Shoulder
      3. Head + Profile (tight crop)
      4. Head (tight crop)
Evaluation

• Visual speaker index (cont.)
  – Effectiveness varies significantly
    • Worst UI requires twice as much time for search as best UI
  – Best results from most information (perhaps intuitive)
    • Head and shoulder shot, front and profile
  – Remaining results not straightforward to interpret

<table>
<thead>
<tr>
<th>Completion</th>
<th>Duration</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>97%</td>
<td>86.72 sec</td>
<td>31</td>
</tr>
<tr>
<td>97%</td>
<td>126.12 sec</td>
<td>30</td>
</tr>
<tr>
<td>91%</td>
<td>137.12 sec</td>
<td>35</td>
</tr>
<tr>
<td>91%</td>
<td>155.94 sec</td>
<td>33</td>
</tr>
</tbody>
</table>
Evaluation

- Some qualitative results
  - Student perspective:
    - Great response to particular UI elements
      - Visual alignment of thumbnails/timeline to text
      - Face index (entertainment factor)
      - Adjustability of parameters
    - Ready availability of video material appreciated
  - Instructor perspective:
    - Benefit from ready availability
      - One-on-one student feedback with A/V material
    - Inclusion in course curriculum
      - Steady improvement in student presentation performance
      - Archive data prepares for first presentation (midterm)
      - Midterm data prepares for final presentation
Overview

- Introduction
  - Motivation
  - Background
  - Challenges
  - Objectives

- Indexing
  - Visual Segmentation
  - Speaker Segmentation
  - Text Index

- User Interface

- Evaluation

- Conclusion / Future Work
Conclusion / Future Work

- Pedagogic requirement for feedback material to improve presentation skills
- Tools must be transparent and inexpensive
- Solution viable, sustainable, and effective
- Areas of improvement:
  - Visual speaker index well-received: need automation
  - Search too simple; need approach for ranking results
  - Use of other approaches for search unexplored
    - What are appropriate semantic concepts for this media?
  - User annotations for better feedback communication
  - Other cues: detect speech flaws (interjections)?
Thank you!

Questions / Answers?

Please note: On-line demo with neutral videos in the works. We cannot disclose current database of videos - our students may get irritated (more than 1500 of them =~ civil warfare). Please contact author for more info on demo. Thanks!