Multimedia II

CSEP 510
Lecture 9, March 1, 2004
Richard Anderson

Announcements

- Lectures
  - Monday, March 1
  - Thursday, March 11
- Homework due dates
  - Thursday, March 4
  - Thursday, March 11

Outline

- Offline use of video
  - Browsing video
  - Video review
  - Video summarization
- Video conferencing
  - Gaze
  - Latency
  - Automatic camera management
- User studies
  - How do you evaluate these systems
  - Evidence that the systems are effective

Offline viewing

- Driving goal
  - Faster viewing
  - Use of video to accomplish some other task
- Observation
  - People are very effective at skimming paper documents

Time compression

- Video speedup
  - Drop a fraction of the frames
  - Increase the display rate
- Audio speedup
  - Lower sampling rate increases pitch
  - Discard segments (33ms every 100ms)
  - Smoothing can improve output signal

Pause removal

- Remove audio and video corresponding to gaps in speech

Compression performance

- Speedup of a factor of 2.0 is tolerable
- Training allows even greater speedups
- Most studies show speedups of about 1.4 when viewers have the choice
- Word rate may be the limiting factor

How do people browse video?

- What techniques to people use to browse video?
- Give them a viewer with additional functionality and see how they use it

Video browsing behavior

- Basic
  - Play
  - Pause
  - Fast-forward
  - Seek
- Enhanced
  - Speed up:
    - Time compression
    - Pause removal
  - Textual indices
    - TOC, notes
  - Visual indices
    - Shot boundary
    - Timeline
  - Jump controls

MSR Video Skimmer

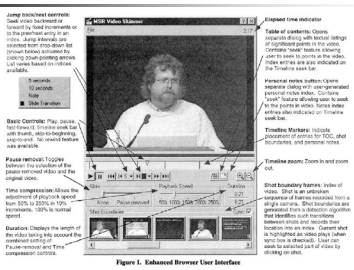


Figure 1. Enhanced Browser User Interface

Study methodology

- Observe participants viewing behavior
  - View video under time constraint
    - 30 minutes for 45-60 minute video
  - Scenario given based on video type
- First with basic browser
- Then twice with enhanced browser

Scenarios

- Classroom
  - Review lecture before a test
- Conference
  - Summarize conference talk for co-workers
- Sports
  - Find highlights in a baseball video
- TV Shows
  - Review missed show before watching final episode of series
- News
  - Summarize news show to family
- Travel
  - Identify interesting segments in a travel video

Results

- 5 viewers per scenario
- Survey to rank features
- Measure number of operations used
- Determine percentage of videos watched

Results

- Different behavior on basic and enhanced
  - Increased viewing percentage
  - Did not use seek / fast forward
- Substantial differences based on scenario
  - Information audio-centric
    - Classrooom, Conference
  - Information video-centric
    - Sports, Travel
  - Entertainment
    - Speedup not desirable

Homework assignment

- Browse a group of videos
- Write outlines
- Vary time available for videos
- You will need a partner for this assignment (but will be able to work by email)

Audio-Video Summarization

- Create a summary video with greatly reduced length
- Domain
  - Informational talks
  - Low production cost

Information Channels

- Audio
- Video
- User Actions
- End user actions
- Slide content

Summary goals

- Conciseness
  - Segments as short as possible
- Coverage
  - All key points covered
- Context
  - Prior segments should establish proper context
- Coherence
  - Segments should flow together

Algorithms

- Given an a video of length t, find a collection of segments S = $\{s_1,\ldots,s_k\}$ such that the total length of S is t' and S is a good summary
- Slide Transition based
- Pitch based
- Use based (combined with slide and pitch)
- Manual (Author based)

Author based

- Author given a text transcript
- Author marked summary segments with a pen
- Author also generated a set of quiz questions for later evaluation

Slide transition based

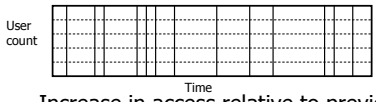
- Show every slide
- Assume content at start of the slide is most important
- Allocated time to slide proportionately to actual time
- Adjust time to allow completed phrases

Pitch based segmentation

- Higher pitch corresponds to more important speech
  - Divide into 1 ms frames
  - Compute pitch for each frame
  - Threshold value: top 1%
  - Each 1 sec window counts number of high pitch frames
  - Divide into 15 second windows
  - Sort by combined score
  - Combine the 15 second windows until total segment length is reached

User access information

- Complete logs of user access
- Typical access



- Increase in access relative to previous slide indicates importance
- Fast drop in access indicates non-importance

Slide, User, Pitch algorithm

- User information to identify more important slides
- Divide slides into thirds based on interest level heuristic
- Slides in first group get 2/3 time, slides in second group get 1/3 time
- Divide slide time inside group based on time watched
- Choose segments per slide based on pitch heuristic

User study

- For informational talks summarized with all four approaches
  - UI Design, IE 5.0, Dynamic HTML, and MS Transaction Server
- 24 subjects from a large software company
  - Subjects received one (1) free espresso drink
- Background test and survey
- Each subject watched all four videos with different summarizations
- After each summary, participants took a quiz and filled out a survey

Results

- Quiz results (before / after)
  - A (2, 5.7)
  - SUP, P, S (2, 4.2)
  - Significant at the .01 level
  - However improvement with auto summarization
- Survey data
  - Significant preference for automatic
  - But SUP, P, S received favorable evaluations
  - Subjects were generally surprised to learn that three of the summaries were automatic
  - Participants evaluation of the later summaries was higher than for the earlier summaries

Follow on study

- Summarization without audio and video
  - Study should have been done first (!)
- Are textual or slide summaries as good as video?
- Same content as previous study

Non-video summaries

- Slides only (SO)
- Text transcript with slides (T)
  - Human transcription used
- Highlighted Transcript with slides (TH)
  - Expert highlights the transcript from above

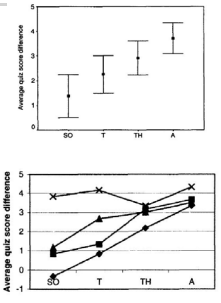
Methodology

- Same as previous study
- Authors had created a group of questions
- Study
  - Pre-test
  - For each video
    - View summary on-line
    - Fill out survey and take quiz

Results

Table 1: Information associated with each presentation.

	UI	DH	IE	MT
Duration (mm:ss)	71:59	40:32	47:01	71:03
# of slides in talk	17	18	27	52
# of slides / min	0.2	0.4	0.6	0.7
# of words in transcript	15229	8061	6760	11578
# of pages in transcript	15	10	8	15
Highlighted words	19%	24%	25%	20%
Duaration of AV summary (mm:ss)	13:44	9:59	11:37	14:20



Survey results

		Synopsis	Key points (%)	Skip talk	Concise	Coherent
Old Study	SPU	4.92	64.17	3.54	4.63	3.58
	P	4.83	62.50	3.04	4.13	3.46
	S	4.33	56.25	3.21	4.08	3.57
	A	5.00	76.25	4.96	5.63	5.33
Current Study	A	4.96	68.91	4.41	5.13	4.13
	TH	4.70	64.13	4.52	4.52	4.35
	T	3.58	61.67	3.83	3.50	4.17
	SO	3.13	41.25	1.96	2.92	2.83

Study Conclusions

- Text transcript with highlighting is competitive with Audio-Video summary
- Top two methods required the most expert effort
  - Continued research in text recognition and text summarization

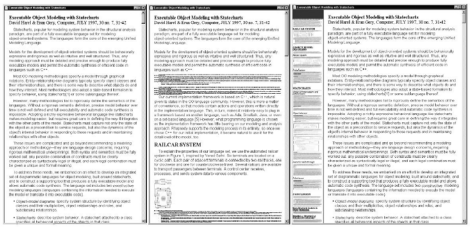
Digression: Reading electronic documents

- Paper reference
- Presenting electronic documents for reading
  - Presentation format
  - Evaluation
- Extracting information
- Evaluation with testing

Document reading

- Scenario
  - Read to learn
  - Read to do
- Layout approaches
  - Linear
  - Fisheye
  - Overview + detail

Layouts



Linear Fisheye Overview + Detail

Experiment

- Evaluate subjects ability to perform tasks based upon reading
- Write essay, answer questions afterwards
  - Essay quality
  - Incidental learning questions
- Direct question answer from papers

Results

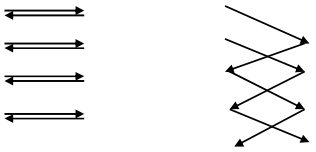
- O+D had significantly better essay scores than L and F
- L and O+D had significantly better incidental learning scores than F
- No significant differences in question answering
- Subjects has a significant preference for O+D
- Efficiency
  - Essay significantly faster using F than O+D or L
  - Question answering significantly faster using L then O+D

Video conferencing issues

- Audio often carries more information than video
- Often harder to get audio right (especially for group video conferencing)
- Processing / bandwidth substantially greater for video than audio
- Tradeoffs
  - Bandwidth vs. Quality
  - Latency vs. Quality
  - Bandwidth vs. Latency

Impact of latency

- Watching the colloquia (or the Oscars)
  - Minimal
- Participating in a video conference



Audio video synchronization

- Audio latency can be lower
  - Coding is more efficient
  - Just use the telephone!
- How close does audio need to be to video to be perceived as synchronized?
- Lip synchronization
  - Talking appears synchronized with lips

Experimental results

- Dixon and Spitz
  - Altered synchronization of video for subject reading prose
  - Subjects pressed but when it appeared out of sync
  - Audio 260 ms behind video or Audio 130 ms ahead of video before being detected
- Steinmetz
  - News reading
  - Shifts of 80 ms not detected
  - Shifts of 160 ms almost always detected
- Miner and Caudell
  - Delays of 200 ms perceived as synchronized
- Television standards – National Association of Broadcasters
  - Audio at most 25 ms ahead
  - Audio at most 40 ms behind

McGurk effect

- Brain perceives conflicting audio and visual as something new
  - Sound "ba" paired with lip movement "ga", people hear "da"
  - Visual stimulus impacts audio with time shift of 200ms
  - Multiple experiments have confirmed this across Western European languages

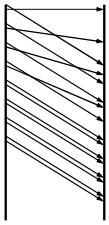
Speech understanding experiments

- Koenig: Understanding of filtered speech impaired with delay of 240ms
- Campbell: Audio masked with white noise. Subjects asked to repeat words. Delay of 400ms (and higher) had significant impact.
- Pandey. Audio masked with multi-talker babble. Delay up to 120 ms comparable to in-sync. Over 120 ms was worse.
- Knoche. Subjects given four syllable non-sense words masked with white noise. Accuracy decreased sharply at 120 ms.

Lip Synchronization Algorithm

- Milton Chen, Stanford
- Assume video has a fixed latency L
- Latency only matters on speaker change
- When speaker starts talking, audio has zero latency. This is gradually increased by stretching audio until it has latency L
  - Audio stretching at start of speech is not detected
  - Latency is reduced in communication rounds

Latency



Intriguing Idea

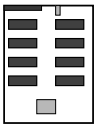
"The perceived round trip audio latency of our algorithm can be equal to the round-trip latency of unsynchronized audio if we can predict the moment an utterance will end."

Gaze

- Vast psychological literature on Gaze
- Gaze important both for direct cues and social value
- Many speculate that the "gaze problem" is a major factor in video conferencing having limited success

Gaze asymmetry

- Look at audience vs. look into camera
- Room setup is the problem in PMP
- Camera placement is critical for desktop video conferencing



Proposed Solutions

- Camera in screen
  - Ideal camera location is in the image!
- Video morphing
  - Software correction of eye positioning
- Making the problem harder – multisite video conference
  - Supporting both look at, and look away

Automatic camera management

- Instructor walks into the room
- Instructor presses the start button
  - Audio, video, recording all start at once
- Instructor delivers lecture
- Instructor presses the stop button
  - Audio, video ends, automatic export of archived material

Lecture room environment

- Capture of lectures
  - Must be inexpensive
  - People cost is dominant, hardware costs have dropped dramatically
  - Primary goal is to capture lectures that weren't previously captured, as opposed to replacing camera operators

Tracking-management problem

- Cameras on lecturer
  - Close shot
  - Long shot
  - Lecturer may move from podium to screen
- Audience camera
  - Occasionally intersperse audience shots
  - Focus on audience members who are talking

Tracking technologies

- Sensor based
  - Accurate but obtrusive
- Vision based
  - Less accurate and can be fooled
- Microphone arrays for locating audience members who are speaking

Video production rules

- Basic goal
  - Automatically produce video that conveys lecture information and is interesting to watch
  - Produce a video that looks like it was done by a human
  - Pass the Turing test

Production rules

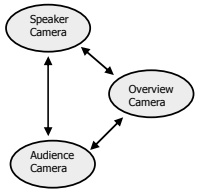
- Framing the speaker
  - Allow sufficient space above speaker's head
  - Don't move speaker tracking camera too often
- Editing rules
  - Establish a first shot
  - Transition to shots that are significantly different
  - Minimum shot durations
  - Maximum shot duration (dependent on camera)
  - Promptly show audience member asking questions
  - Occasionally show audience when no questions arise

Camera tracking

- Lecturer camera
  - Track the speaker, enter not ready state when speaker is lost
- Audience camera
  - Focus on audience member who is speaking
  - Revert to general position when no one in audience is speaking

Virtual director

- Finite state machine
- State change events
  - Status change
    - Ready, not ready
  - Time expire



Evaluation

- Comparison of automatic system with human controlled system
- Film same lecture with both systems
- Have people watch both systems and answer questionnaires
  - Field study (on desktop machines)
  - Lab study (under supervision, so subjects weren't reading their email!)
- Results positive
  - Subjects had difficulty telling which was automatic and which was manual
  - Many questions were hard to answer, because people are not aware of them when watching video

Evaluation by experts

- A second study was done after significant refinement of capture system
- A series of lectures was filmed by system and professional videographers
- Evaluation by videographers and subjects

Survey questions	Profess. evaluate system	Audience evaluate system	Audience evaluate profess.
1. Shot change frequency	2.5 (2.8)	3.0 (2.6)	4.0 (3.4)
2. Framed shots well	1.5 (1.8)	3.0 (2.7)	4.0 (3.6)
3. Followed lecturer smoothly	2.0 (2.0)	2.0 (2.3)	4.0 (3.5)
4. Showed audience questioner	3.5 (3.5)	3.0 (2.8)	2.0 (2.7)
5. Showed audience reaction	4.0 (3.5)	2.0 (2.3)	2.0 (2.3)
6. Showed facial expression	3.0 (2.8)	2.5 (2.8)	3.0 (3.2)
7. Showed gestures	2.5 (3.2)	4.0 (3.2)	4.0 (3.5)
8. Showed what I wanted to watch	3.0 (3.2)	4.0 (3.4)	4.0 (3.9)
9. Overall quality	2.0 (2.0)	3.0 (2.8)	4.0 (3.8)
10. As compared with previous experience	1.5 (1.5)	3.0 (3.1)	3.0 (3.6)

Problem shots

Figure 4. Examples of bad framing. (a). Not centered. (b). Inclusion of the screen edge. (c). Too much headroom. (d). Showing an almost empty audience shot.

Detailed rules

- Study suggested many production rules
- Rules evaluated for technical feasibility in an automated system

Tracking and framing rules

- 2.1 Keep a tight head shot
- 2.2 Center the lecturer but balance for lecturers gaze or gesture
- 2.3 Track lecturer smoothly
- 2.4 Track lecturer or switch cameras depends on context

Audience rules

- Promptly show audience questioners
- Avoid empty audience shots
- Occasional show the audience when there are no questions

Shot transitions

- 4.1 Reasonably frequent shot changes
- 4.3 Maximum duration depends on type
- 4.4 Shot transitions should be motivated
- 4.6 Overview shot is a good backup

Expert advice summary

- Validation of system
  - "It did exactly what it was supposed to do … it documented the lecturer, it went to the questioner when there was a question"
- Very different evaluation from average viewers
  - Sensitive to different issues
- Very rich set of rules derived
  - Some could be implemented easily, others very hard

Lecture summary

- Video browsing
  - Compression
  - Skimming
  - Summarization
- Summarization
  - Video
  - Separate media
  - Reading
- Video
  - Latency
  - Gaze
- Automatic camera management
  - User evaluation
  - Expert evaluation
- User studies