Video Summarisation

• Summary == condensed version of something so that judgments about the full thing can be made in less time and effort than using the full thing

• Summaries have widespread application as surrogates resulting from searches, as previews, as familiarisation with unknown collections

• Video summaries can be keyframes (static storyboards, dynamic slideshows), skims (fixed or variable speed) or multi-dimensional browsers

• Literature & previous work shows interest in evaluating summaries, but datasets always small, single-site, closed
Summarisation Data

- 42 files as development data, 40 files as test data (- one withdrawn)
- Mostly scripted dialogue, environmental sounds, much repeating (==redundancy), wasted shots, clapboards and colourbars

Test video durations (min)

- Test videos - mean duration: 26.6 min (max: 36.9 min.; min 9.8 min.)
- Example of full one full rushes video MS221050
System task

• Create an MPEG-1 summary of each file
• Each summary <= 2% of the original
  – twice as compact as in 2007
• Dual evaluation criteria were to
  – Eliminate redundancy
  – Maximise viewers’ efficiency at recognising objects & events as quickly as possible
• Interaction limited to:
  – Single playback via mplayer in 125 mm x 102 mm window at 25 fps with unlimited optional pauses
How to evaluate the rushes summaries?

• Seems intractable in the general case:
  – Formally identify all the content of an original video
  – Do likewise for a summary, and then
  – Compare them, in a way which is repeatable and affordable

• So we approximated for the data at hand:
  – Humans created partial ground truth for the original (40) videos
    • Identify important segments using any distinctive object/event
    • Accept variability due to differences in human judgment
  – Human viewed each summary and judged it against the list of important segments (ground truth)
Sample ground truth (MS221050)

- 2 men in white carry man in hooded blue shirt
- Head and shoulders of red-headed woman visible
- Close up of red-headed woman (head and neck only visible)
- Red-headed woman & man in leather jacket (waist up visible) stand while man in white enters.
- Man in blue shirt and man in suit stand and talk, head and shoulders of both visible
- Man with purple shirt and man in blue shirt stand and talk, head and shoulders of both visible
- Man in white coat seated, waist up, side view
- Close up (head visible) of black man
- Close up (head visible) of black man with blue wrap on shoulders
- Group of people walking toward camera carrying large chest
- Group of people sitting around desk
- 3 people enter and stand left of desk
- Man and woman seated, face camera, head and shoulders visible
- Man in blue shirt and red-headed woman stand (head and upper chest visible)
(approx. 30min)

Rushes 1
Rushes 2
Rushes 3
Rushes 39

39 Rushes video

Ground Truth Items
(24 / Rushes)

1. A man standing beside...
2. :
3. :

39 GT Items lists

Randomly selected GT Items (12 / Rushes)

1. A man standing beside...
2. :

39 GT Item lists

10 assessors

All summaries of Rushes 1 grouped, randomized, all assigned to 3 Assessors...

Measures

$\text{Fraction of GT included}$

$\text{Lack of junk video}$

$\text{Lack of redundancy}$

$\text{Pleasant tempo and rhythm}$

$\text{GT Assessment time}$

Total 39 x 31 = 1209 Video Summaries submitted

Total 31 Participants

Participant 1

Summary 1
Summary 2
Summary 3
Summary 39

Participant 2

Summary 1
Summary 2
Summary 3
Summary 39

Participant 31

Summary 1
Summary 2
Summary 3
Summary 39

... after watching original Rushes 1 video and examining its GT Items

2%
Measures

• Subjective:
  – Fraction of (up to 12 items of) ground truth found
  – Lack of junk (color bars, clapboards, all white/black frames)
  – Pleasant tempo and rhythm
  – Lack of redundant video

• Objective:
  – Assessment time to judge included ground truth
  – Summary duration
  – Summary creation compute time

• Additional data:
  – Number/duration of pauses in assessment of included segments
  – Feedback on assessment software, procedure, experience
Participating groups’ approaches

- 26 of 31 groups had papers at the TVS’08 ACM MM workshop so we know a bit about their approaches - though no structured description
- Most groups, almost all, explicitly searched for and removed junk frames;
- Most groups, majority, used some form of clustering of shots/scenes in order to detect redundancy;
- Several groups included face detection as some component;
- Most groups used visual-only, though some also used audio in selecting segments to include in summary;
- Camera motion/optical flow was used by some groups;
- Finally, most groups used whole frame for selecting, though some also used frame regions;
Summary generation

• There was much more variety among techniques for summary generation than among techniques in summary selection;
• Many groups used FF or VS/FF video playback;
• Several groups incorporated visual indicator(s) of offset into original video source, within the summary;
• Some used an overall storyboard of keyframes;
• Some used keyframe playback but most used the unaltered original video, perhaps using sub-shots only;
• Some used non-hard cut shot transitions, and one did progressive summary generation, on-the-fly;
Results: fraction GT included

Medians: 0.08 – 0.83
Baseline significantly better than all other runs
In general, FF approach seems to correlate with higher scores.

Looking at included GT per unit summary duration:
Asahikasei.1 does better than baseline
Results: lack of junk

Medians: 2.33 – 3.67

Baseline drops to bottom – as expected if the evaluation is working, since baseline makes no attempt to remove junk, just to move it past the viewer faster

Most scores in a narrow range

Bottom systems are all and only the FF systems???
Results: lack of redundancy

Medians: 2 - 4

Again, baseline drops to bottom as expected (no attempt to remove redundancy)

Most scores in an even narrower range than “lack of junk”
Results: pleasant tempo

Medians: 1.33 – 3.33
Wider range at low end
Using FF correlates with low scores on pleasant rhythm?
Results: assessment time

Medians: 21.67 – 61.67 (s)

Seems more time spent judging again correlated with higher inclusion scores .. But which was cause and which was effect?
Results: summary duration

Almost all smaller than target
No penalty, no reward in the measures
Longer summaries don’t imply more ground truth included
Results: summary creation time

Median times just under 20 minutes
Some very fast
Some very expensive (unoptimized for time, e.g. IRIM genetic algorithm)
Evaluating the evaluation

• No problems in ground truth creation or assessment

• Agreement in binary judgments of included ground truth
good again 81% (versus 78% in 2007; 50% expected by chance)
  – Fraction of agreement on a judgment of “no inclusion” was 53.8%
    (about the same as in 2007 (57.2%))

• Pairwise differences in well-formedness judgments
smaller than in 2007
  – 2008 mean and median differences: ~ 1.0
  – 2007 mean differences:
    • 1.442 for ease of understanding
    • 1.366 for redundancy
Final observations

- Evaluation framework passes sanity checks again
- Systems achieved compression target of 2%, moving from 4% in 2007 – let’s not underestimate this challenge
- Use of fast forward spread to ~ 50% of runs
- Baseline really only aimed to include ground truth – not a baseline for well-formedness
  - very high on included ground truth
  - very low on usability measures
- Computation time to generate summaries varied wildly
- Is this problem now solved?
- What should summarisation move on to next?
Thanks to ...

• BBC Archives and Richard Wright
• NIST and Intelligence Advanced Research Projects Activity (IARPA)
• European Commission under contract FP6-027026 (K-Space)
• The assessors at NIST who created the ground truth and the assessors at Dublin City University for the evaluation
• Philip Kelly at Dublin City University for helping to organize the judging
• Carnegie Mellon University for providing the baseline results once again
• Several sites for mirroring the video data
• The program committee and others for reviewing papers
• All the participating groups for taking part
Possible continuations… *mobisodes*

• More BBC rushes video is available, but
  – Systems are doing well on the current measures
    • time to see how well real users like the results
  – System approaches are converging

• BBC also interested in automatic summarization of
  – *produced video* for mobile devices (mobisodes)
    • catch-up: find the video in episode x needed to understand episode x+1
    • preview: find the video in an episode that will make a viewer want to see the
      episode but without destroying suspense

• Lots of questions remain:
  – availability of production data beyond video?
    • audio description
    • script
    • closed captioning
  – how to evaluate
    • effectiveness
      – manually describe needed video as was done with rushes?
    • usability (especially for a mobile device … Which? In what setting? By whom?)