TRECVID BBC Rushes Summarization Pilot Workshop

Paul Over

Information Access Division
National Institute of Standards and Technology

Alan F. Smeaton

Philip Kelly

Adaptive Information Cluster
Dublin City University

NIST
National Institute of Standards and Technology

28. September 2007

DCU
DUBLIN CITY UNIVERSITY

Fasttrack timeline

January
- initial ideas at NIST and DCU

February
- discussions with others;
- ground truthing starts

March
- guidelines, development data available

April
- test data available
- ground truth complete
- evaluation software complete

May
- system output submitted for 2 week evaluation at NIST

June
- evaluation results returned;
- initial papers due

July
- final papers due

And somewhere in there systems got developed!

Video Summarisation

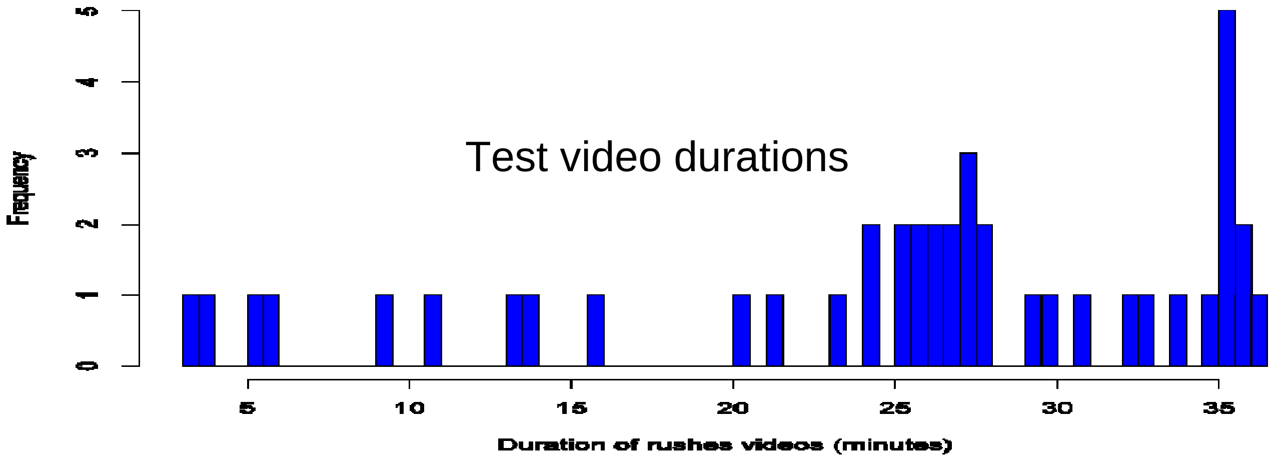
- Summary == condensed version of something so that judgments about the full thing can be made in less time and effort than the full thing
- In a world of information overload, summaries have widespread application as surrogates resulting from searches, as previews, as familiarisation with unknown collections
- 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

- BBC provided 11 boxes of BETA SP tapes, 250 hours of rushes from dramatic series… Casualty, House of Elliot, Jonathan Creek, Ancient Greece, Between the Lines & other miscellaneous

Summarisation Data

- Rushes … we digitised 100 hours, each tape -> 1 file, average 25 minutes



- Took a random sample of each source
- **Mostly scripted dialogue**, environmental sounds, much repeating (==redundancy), wasted shots, clapboards and colourbars
- 50 files as development data (ground truth for 21), 42 files as test data
- Example of full one full rushes video

System task

- Create an MPEG-1 summary of each file
- Each summary <= 4% of the original
- 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 (42) 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 (MRS044500)

Similar action from varying points of view

2 men in dark suits walk past Ford truck to building entrance

2 men in dark suits enter building

person in brown coat opens rear end car and removes wheelchair (seen from front of car)

woman walks around car to passenger window (seen from rear end of car)

close up of man in passenger seat (seen from front of car)

woman in brown coat removes wheelchair and brings it round to the passenger door (seen from front of car)

man in beige suit appears (seen from front of car)

man in beige suit opens car door (seen from front of car)

woman in brown jacket undoes man in car's seatbelt (seen from front of car)

woman in brown jacket helps passenger into wheelchair (seen from front of car)

close up of man in striped top in car viewed from driver seat

man in beige suit appears (seen from inside car)

man in beige suit opens car door (seen from inside car)

woman in brown jacket undoes man in car's seatbelt (seen from inside car)

woman in brown jacket helps passenger into wheelchair (seen from inside car)

man in beige suit appears (seen closeup from front of car)

man in beige suit opens car door (seen closeup from front of car)

woman in brown jacket undoes man in car's seatbelt (seen closeup from front of car)

woman in brown jacket helps passenger into wheelchair (seen closeup from front of car)

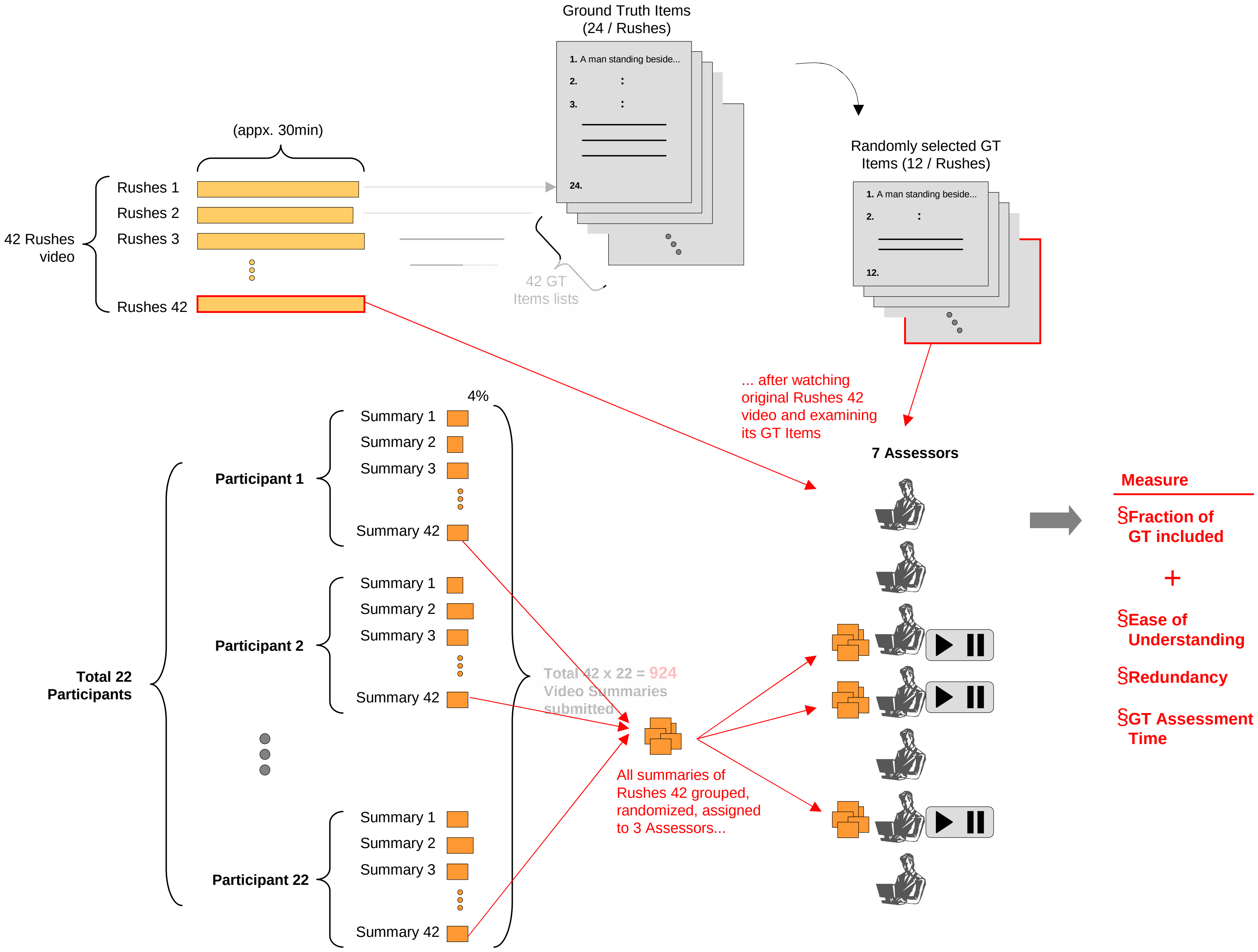
ETC.

Sample ground truth #2

1. closeup of woman, in shadow, rubbing her face
2. man sits on the ground with a monkey on a chair, no one else in the room
3. man sits on the ground with a monkey on a chair, old woman enters scene
4. man sits on the ground with a monkey on a chair, old woman stands at table
5. man sits on the ground with a monkey on a chair, old woman stands at table, woman in red enters scene
6. man sits on the ground with a monkey on a chair, two women stands at table
7. man sits on the ground with a monkey on a chair, old woman stands at table, woman in red exits scene
8. man sits on the ground with a monkey on a chair, old woman exits scene
9. zoom in on man sitting on the ground with a monkey on a chair and old woman standing at table
10. man sits on the ground with a monkey on a chair, old woman standing at table, woman in red sitting
11. man sits on the ground with a monkey on a chair, mans legs visible
12. man sits on the ground with a monkey on a chair, mans legs visible, woman passes in from of them
13. man sits on the ground, monkey exits scene, mans legs visible,
14. man sits on the ground stands up and exits scene
15. closeup of monkey on the chair
16. empty chair
17. closeup of mans, head and shoulders only visible
18. closeup of blonde womans face, head and shoulders only visible
19. closeup of blonde womans face, head and shoulders only visible, old woman visible in the background
20. empty kitchen scene, two doors visible in the background
21. old woman enters kitchen scene, two doors visible in the background
22. woman in red enters kitchen scene, two doors visible in the background
23. two women standing at table in kitchen scene, two doors visible in the background
24. woman in red sits down at the table in the kitchen, two doors visible in the background
25. man walks around kitchen, two doors visible in the background
26. man picks up cup off of kitchen table
27. camera pans left as old woman walks through doorway
28. old woman picks up pad of paper off of cabinet
29. camera pans left as woman in red walks through doorway
30. woman in red reads letter in her hand
31. woman in red fixes her hair in a mirror
32. closeup as woman in red sitting down, no one else is visible
33. woman in red sits picks up coffee jug
34. closeup of woman in red, while woman in blue walks behind her
35. closeup as woman in red drinks, while man walks behind her

Assessor assignments

Videos	First assessor	Second assessor	Third assessor
1-6	A	D	E
7-12	A	B	F
13-18	B	C	G
19-24	A	E	F
25-30	C	F	G
31-36	B	D	G
37-42	C	D	E


Ground Truth Items
(24 / Rushes)
1. A man standing beside...
2. :
3. :
24.
Randomly selected GT Items (12 / Rushes)
1. A man standing beside...
2. :
12.
(appx. 30min)
42 Rushes video
Rushes 1
Rushes 2
Rushes 3
Rushes 42
42 GT Items lists
... after watching original Rushes 42 video and examining its GT Items
7 Assessors
Measure
§Fraction of GT included
+
§Ease of Understanding
§Redundancy
§GT Assessment Time
4%
Total 22 Participants
Participant 1
Participant 2
Participant 22
Summary 1
Summary 2
Summary 3
Summary 42
Total 42 x 22 = 924 Video Summaries submitted
All summaries of Rushes 42 grouped, randomized, assigned to 3 Assessors...

Measures

- Subjective:
  - Fraction of (12 items of) ground truth found
  - Ease of use
  - Amount of near-redundancy
- 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

22 Participating groups' approaches

1. At&T: shot clustering to remove redundancy, use shot with most speech/faces;
2. Brno Univ.: cluster shots using PCA, remove junk shots;
3. CMU: k-means clustering using iterative colour matching, audio coherence;
4. City UHK: obj. detection, camera motion, keypoint matching for repetitive shots;
5. Columbia: duplicate shot detection and ASR;
6. Cost292: face, camera motion, audio excitement;
7. Curtin U: shot clustering using SIFT matching;
8. DCU: amount of motion & faces for keyframe selection;
9. FXPAL: colour distribution, camera motion, for repetition detection;
10. HUT: SOMs for shot pruning to eliminate redundancy;
11. HKPU: junk shot removal, visual & aural redundancy;

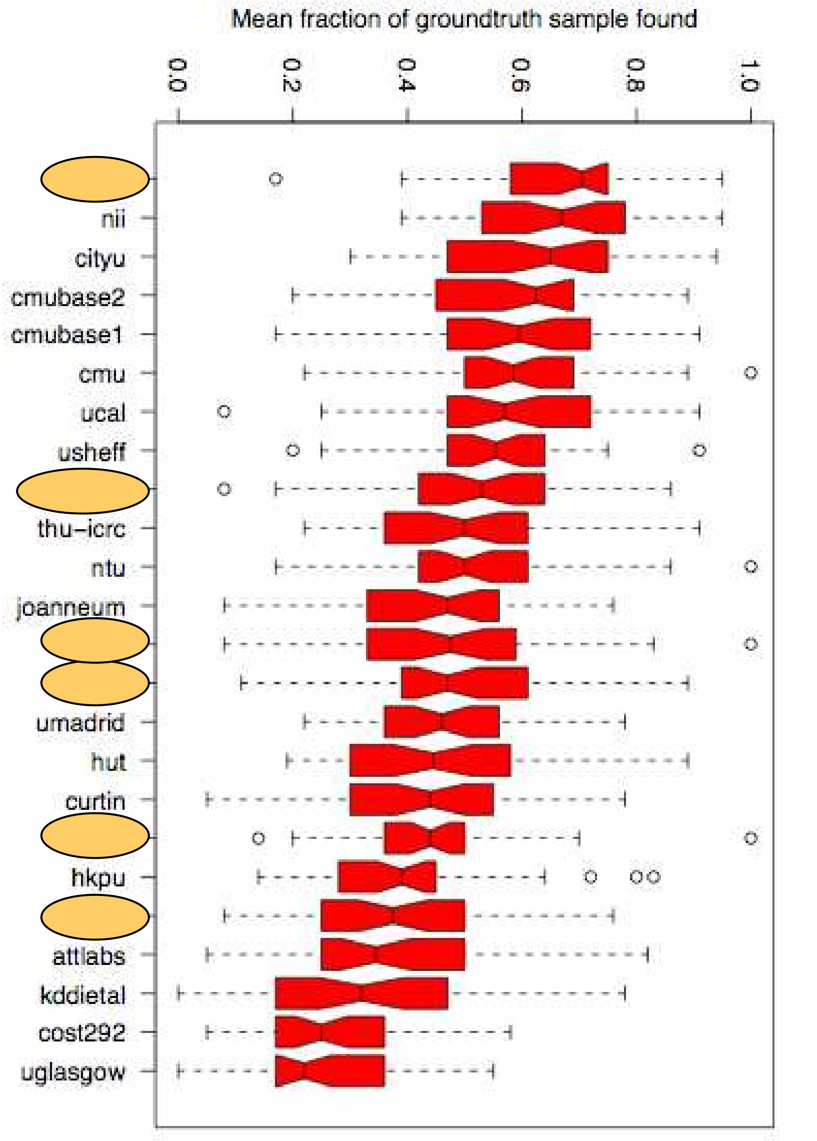
Participant approaches (continued)

12. Eurecom: determine the most non-redundant shots;
13. Joanneum: variant of LCSS to cluster re-takes of same scene;
14. KDDI: use only low-level features for fast summarisation;
15. LIP6: eliminate repeating shots using ‘stacking’ technique;
16. NII: feature extraction and clustering;
17. Natl. Taiwan U: LL shot similarity & motion vectors, then cluster;
18. Tsinghua/Intel: keyframe clustering, repetitive segments, main scenes/actors;
19. UCSB: k-means clustering on HL features, speech, camera motion;
20. Glasgow: 0-1 knapsack optiisation problem, shot clustering;
21. UA Madrid: single pass for realtime clustering on-the-fly, colour-based;
22. Sheffield: concatenate some frames from middle of each shot;

Summary formats

- Plain keyframes: Glasgow
- Plain clips: COST292, Curtin, HKPU, KDDI, Madrid, NTU, Sheffield, CMU baselines
- Clips of 1s duration: CMU, CUHK, Helsinki, UCSB
- Clips FF: NII
- Main scene/actor, then clips: Tsinghua
- Clips with numeric/text indicators of offset/re-takes: AT&T, JRS
  - **Columbia** - clips w. picture in picture showing repetition & also showing numeric offsets
  - **U Brno** - clips w. picture in picture showing iconic scrollbar offsets, redundancy, scrollbar progress
  - **Eurecom** - clips in 4-windows, FF, clustered, no indicators
  - **LIP6** - clips with VSFF, speed indicator and numeric offsets
  - **FXPAL** - clips with variable speed FF and numeric and iconic offset
  - **DCU** - KFs w/ metadata showing offset, faces, motion

Results: fraction GT/ease of use



Range is 0.22 to 0.70

Baselines were good !

cityu, lip6, nii - better than baselines ($p<0.05$)

Disappointing for the non-KF, non-clip interfaces

Results: fraction GT/ease of use

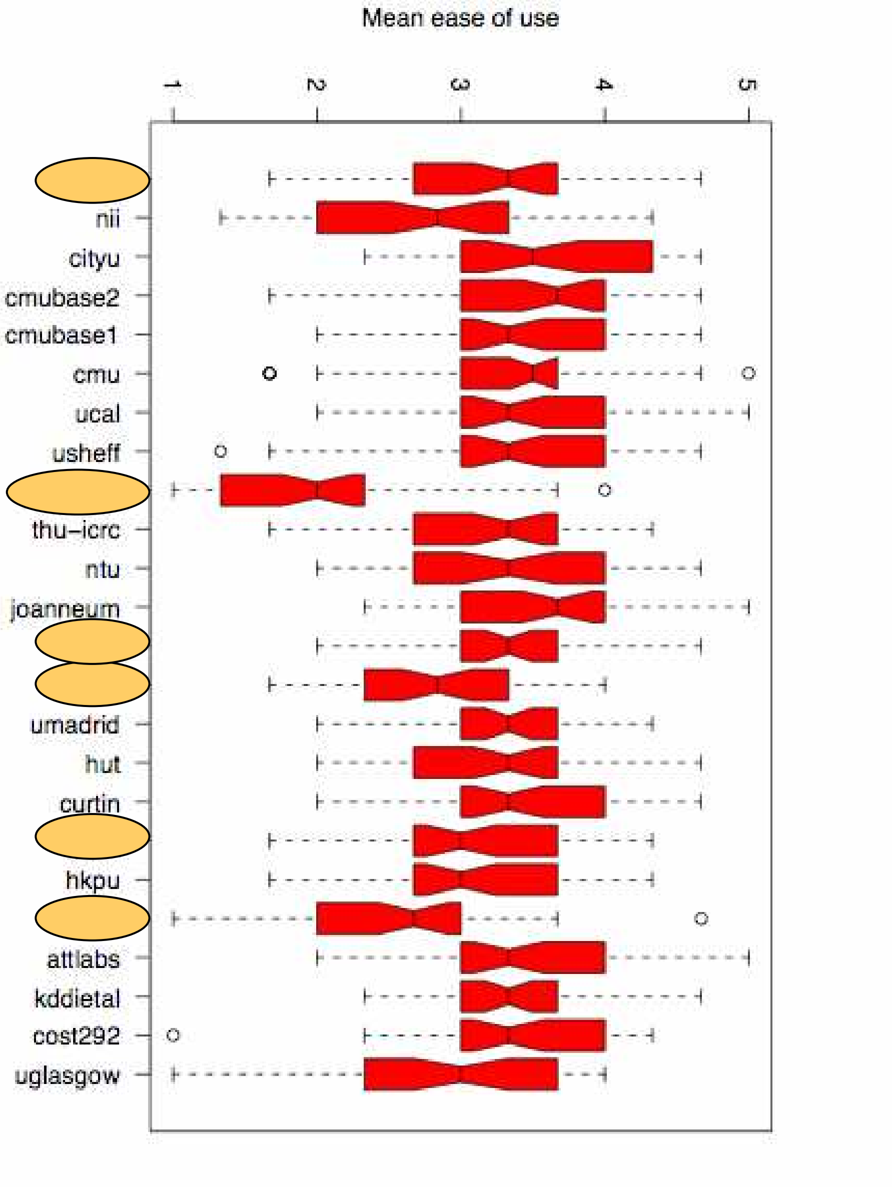
1 (bad) to 5 (good)

Almost all the same, except non-standard interfaces

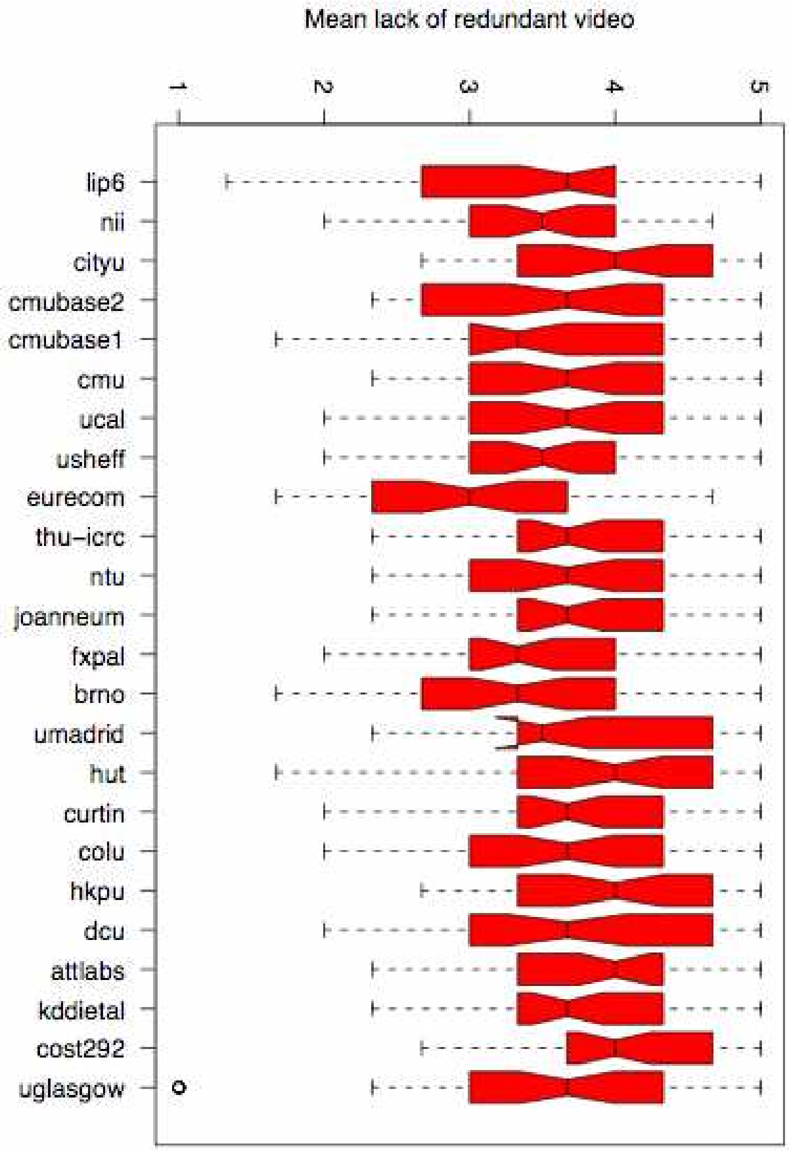
Non-standard presentations got most of the lowest scores

Only cityu better than baselines (p<0.05)

Tells us assessors need training?, interfaces need easier learnability?



Results: lack of redundancy



Most scores ranged narrowly between 3 and 4

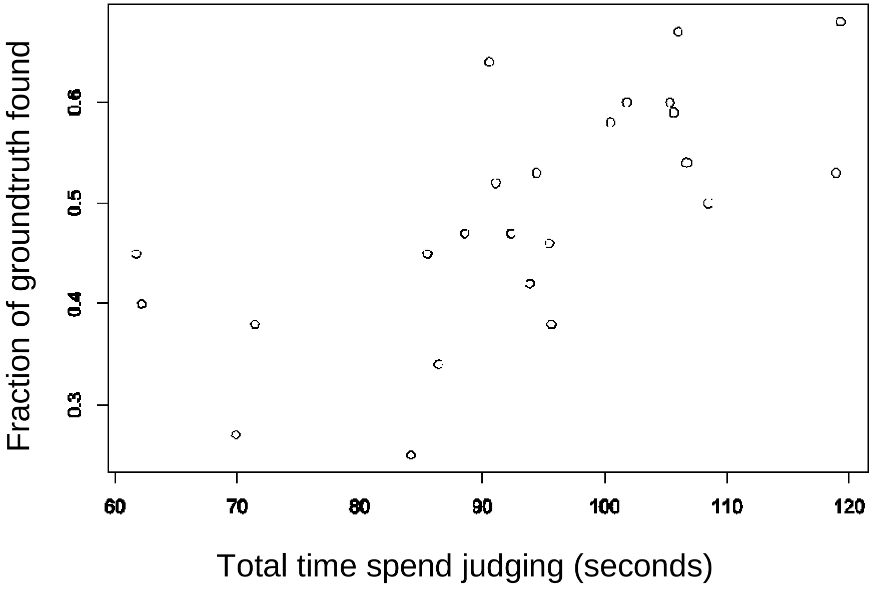
Most systems better than baselines ($p<0.05$)

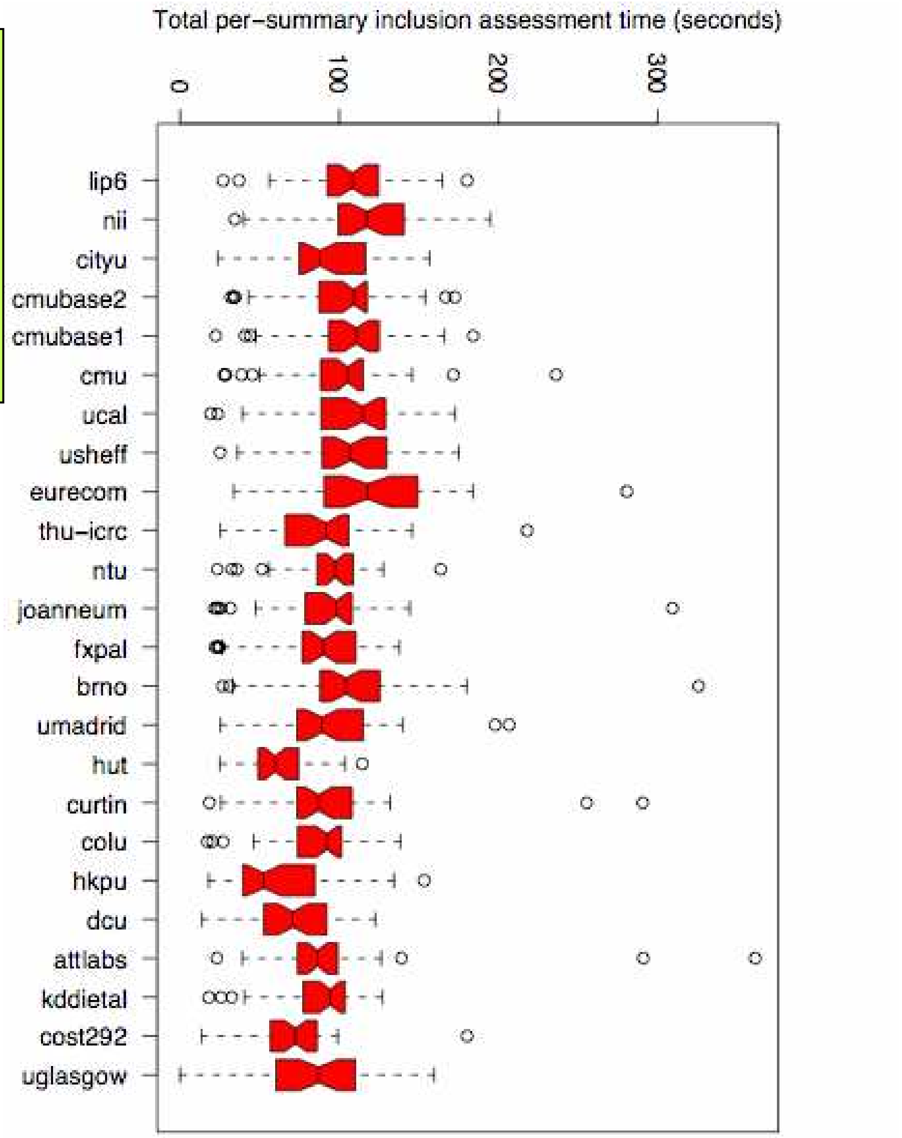
5-point Likert scale too coarse?

Results: assessment time

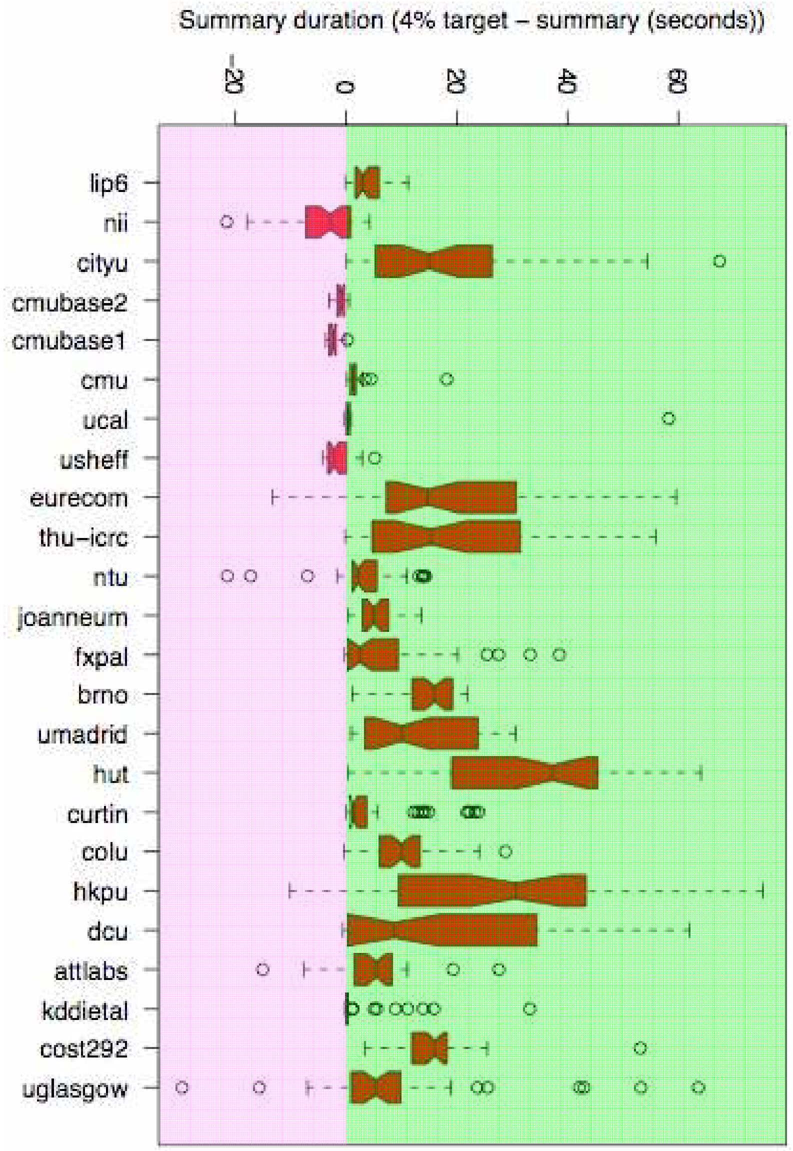
Median times from 53s - 118s

Seems more time spent judging correlated with higher inclusion scores .. But which was cause and which was effect ?





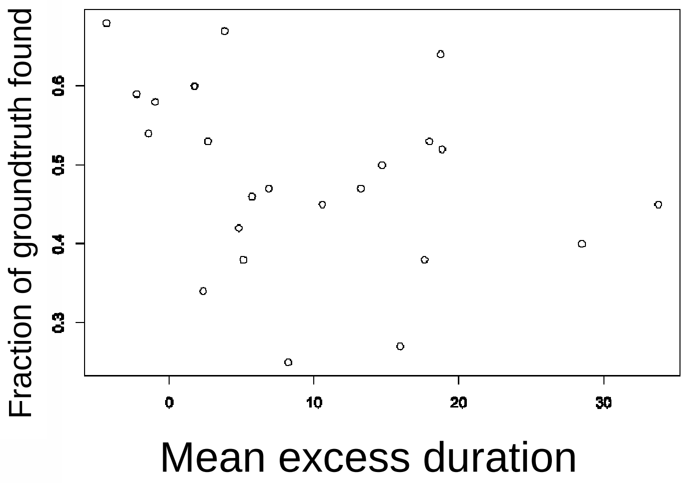
Results: summary duration/creation time



Almost all smaller than target

No penalty, no reward in the measures

Longer summaries don’t imply more groundtruth included

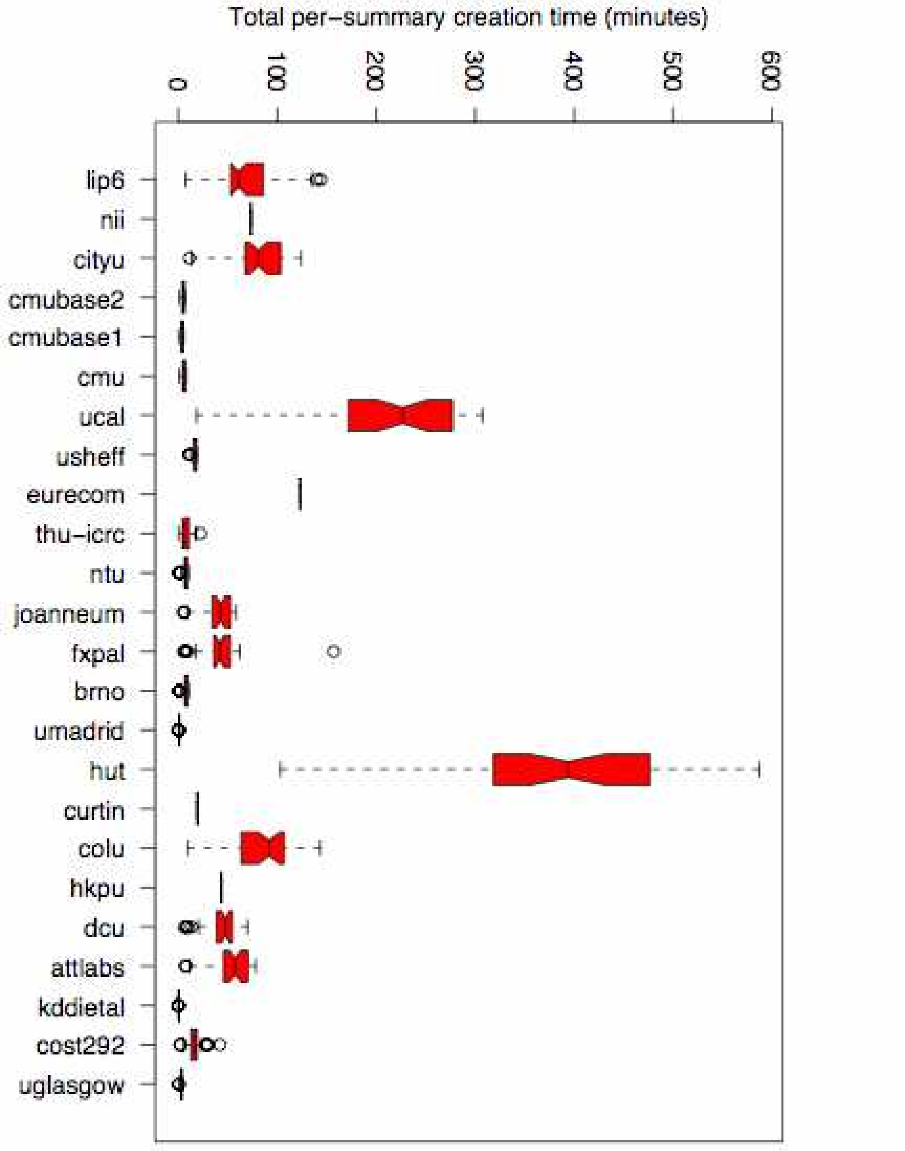


Results: summary creation time

Median times just under 20 minutes

Some very fast

Some very expensive (unoptimized for time?)

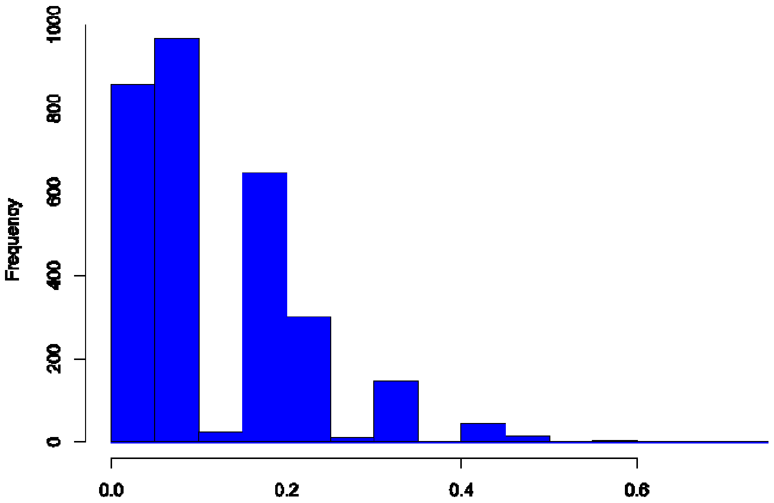


Evaluating the evaluation

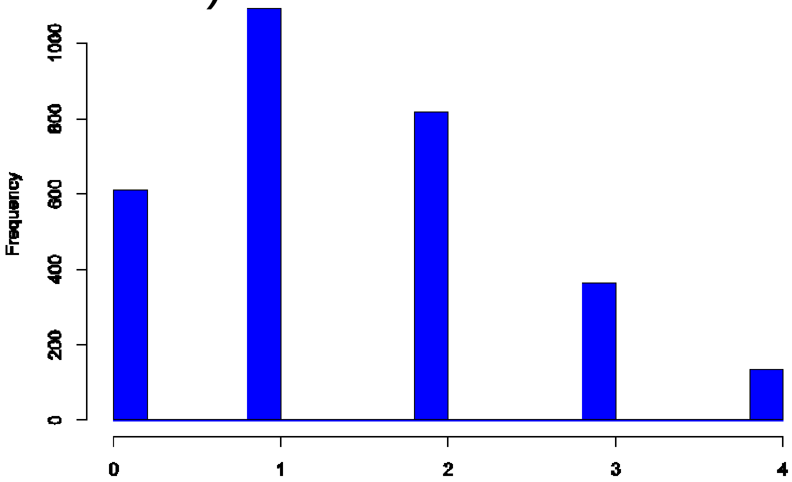
- Ground truth creation
  - A number of issues were identified and addressed so ground truthing was doable for almost all of the videos in the time allowed
  - Ground truth for relatively unstructured, non-repeating videos raised additional issues not addressed
  - The problem of wide variation in granularity and its impacts need further study
- Assessor feedback
  - Learned the procedure/software quickly
  - Expressed confidence in ability to judge included truth correctly
  - Some found use of 5-point scale difficult
  - Some complained about summaries with multiple simultaneous views
  - One indicated the “ease of use” question was misunderstood but results seem to reflect consistent dislike of non-standard presentations

Evaluating the evaluation: score differences

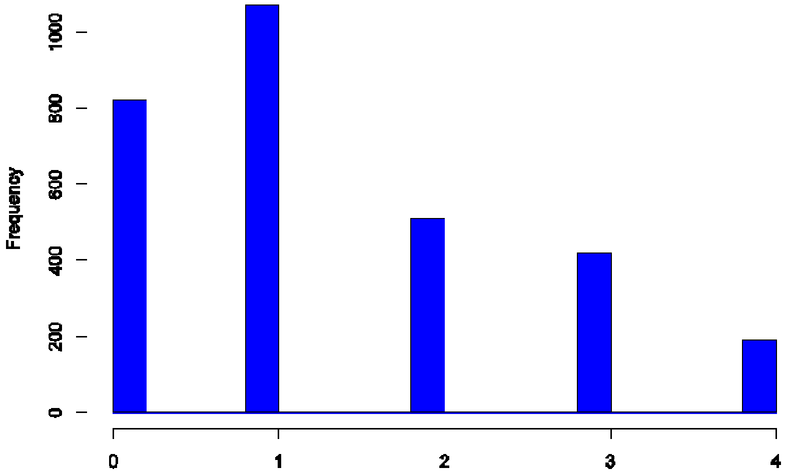
- Pairwise score differences seem low for ground truth found
- Apparently less agreement for “ease of use” and “lack of redundancy”?
- Agreement in binary judgments of included ground truth good: 78% (versus 50% expected by chance)



Pairwise score differences for ground truth found



Pairwise score differences for ease of understanding



Pairwise score differences for lack of redundancy

Evaluating the evaluation: effect of score variability on significant system differences

- Withholding an assessor’s judgments:
  - 7 sets of results – each based on withholding a different assessor’s judgments
  - Randomization test at 0.05 level of significance run on each set of results
  - No significantly different systems swapped positions due to use of different results
- Nearly same results when comparing results produced by 3 random sets of single-judgment assignments to each other or the official results.

Conclusions

- This is the first large-scale, multi-participant evaluation of summarisation of video
- Seems to pass feasibility and sanity of results tests
- 2007 concentrated
  - more on “did the summary contain all the clear and important material in the original”
  - less on issues of redundancy in summary and learnability of summary formats
- Good agreement on the inclusion of GT in summaries, the most detailed component of evaluation and
- 4% target could have been even smaller
- Note: 2007 TRECVid summarisation tied to nature of data - TV series rushes, and techniques not necessarily generalisable to other kinds of rushes or non-rushes