Brief and High-Interest Video Summary Generation

Evaluating the AT&T Labs Rushes Summarizations

Zhu Liu, Eric Zavesky*, Behzad Shahraray, David Gibbon, Andrea Basso

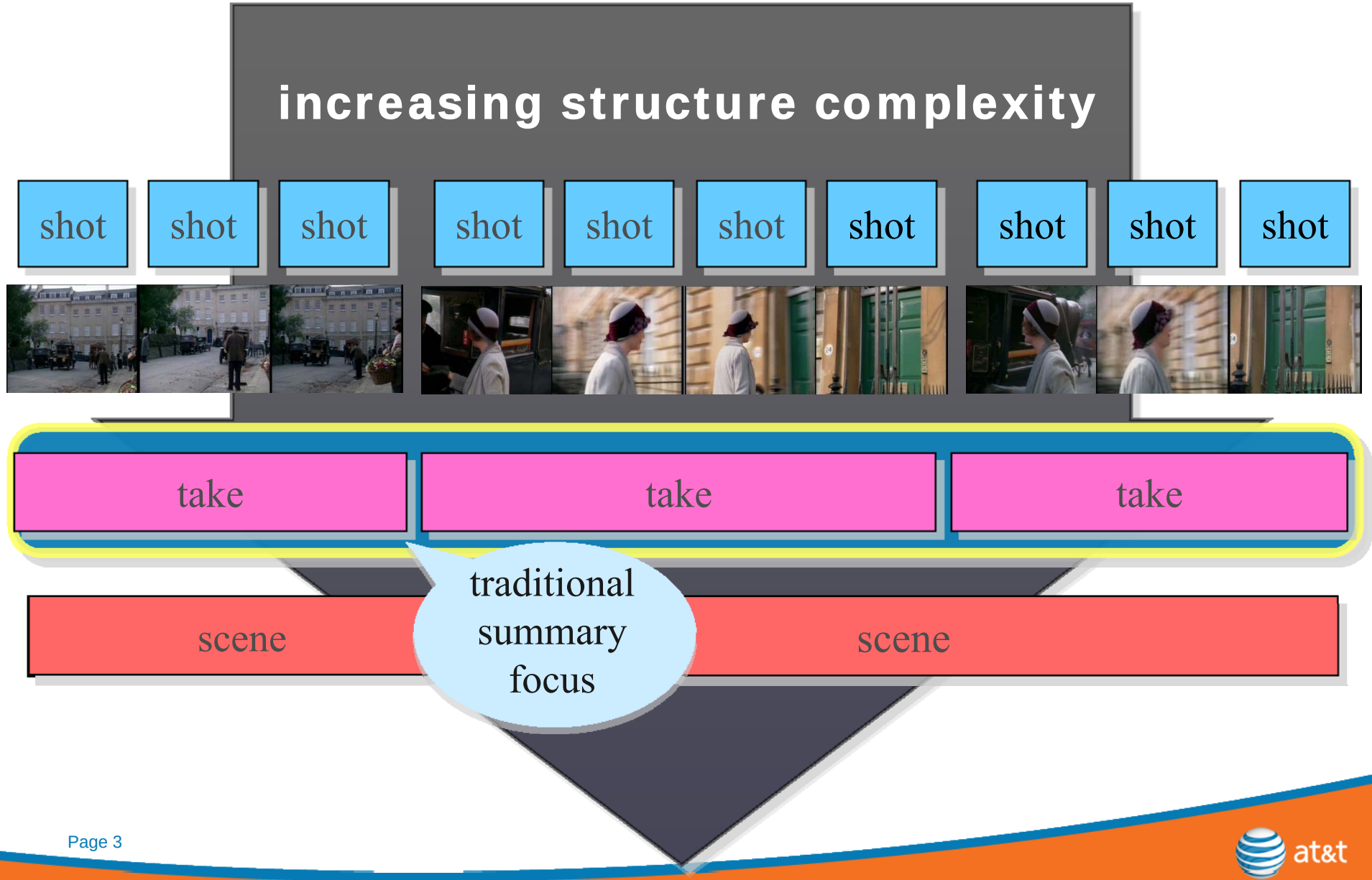
AT&T Labs – Research
***Columbia University**

zliu@research.att.com

Outline

- **Introduction**
  - Structure review
  - Definition of a summary
  - New user-centric goals for this year
- Our approach
- Evaluation results
- Conclusion

BBC Structure Visualization



What does summarization mean?

street scene with man sweeping

Woman walks up steps and goes through green door (seen from sidewalk)

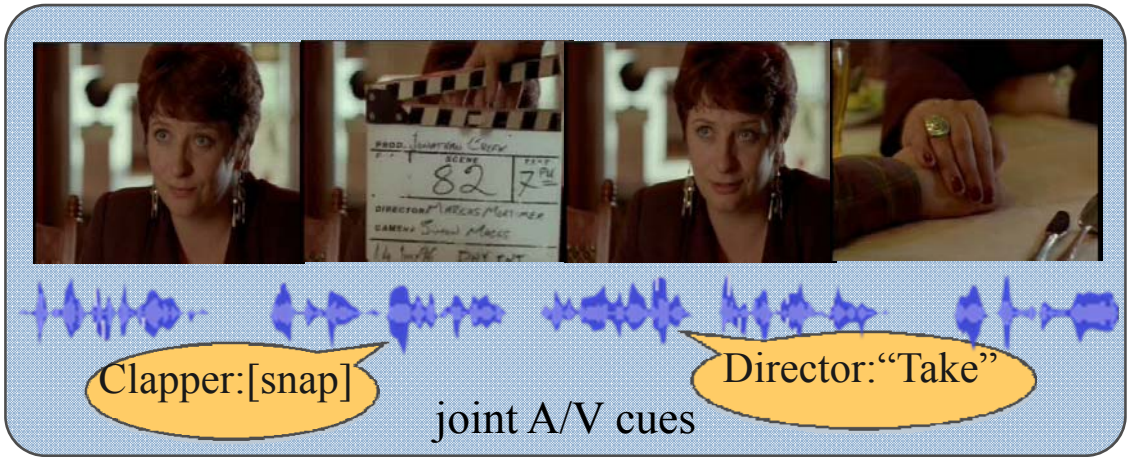
Woman exits taxi and pays driver (seen from sidewalk)

Woman exits taxi and pays driver (seen from street)

Woman walks up steps and goes through green door (seen from street)

MRS045104 ground truth examples

Goal: User Centric Summarization

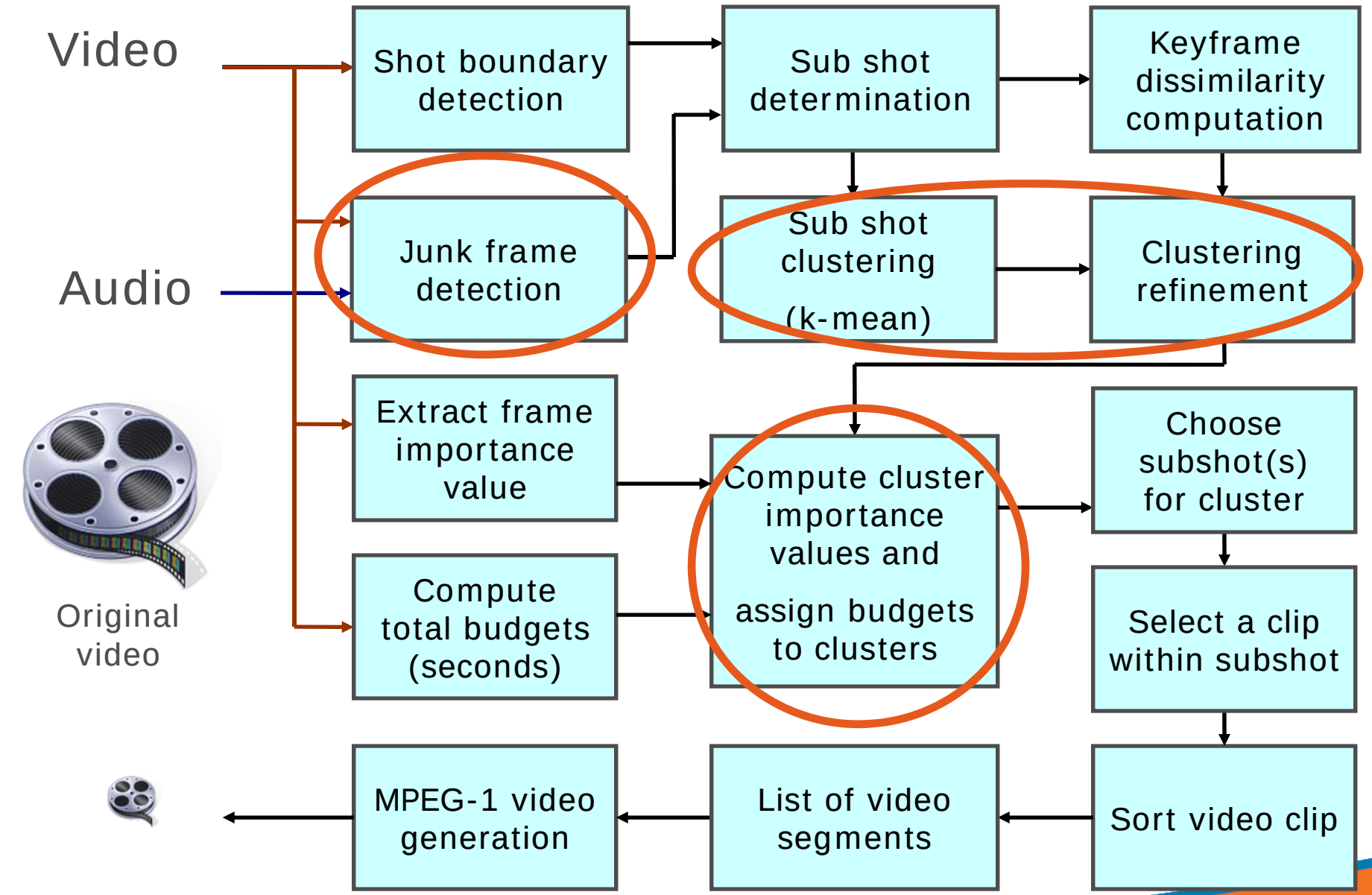


1. Empirically, cluster target size can be tuned to number of takes (~3-4)
2. Audio-visual features can be used together to detect relevant content
3. Choose shots for summary based on user-centric model assisted by low-level features
4. Intuitive, non-cluttered summary display with intelligible video segments

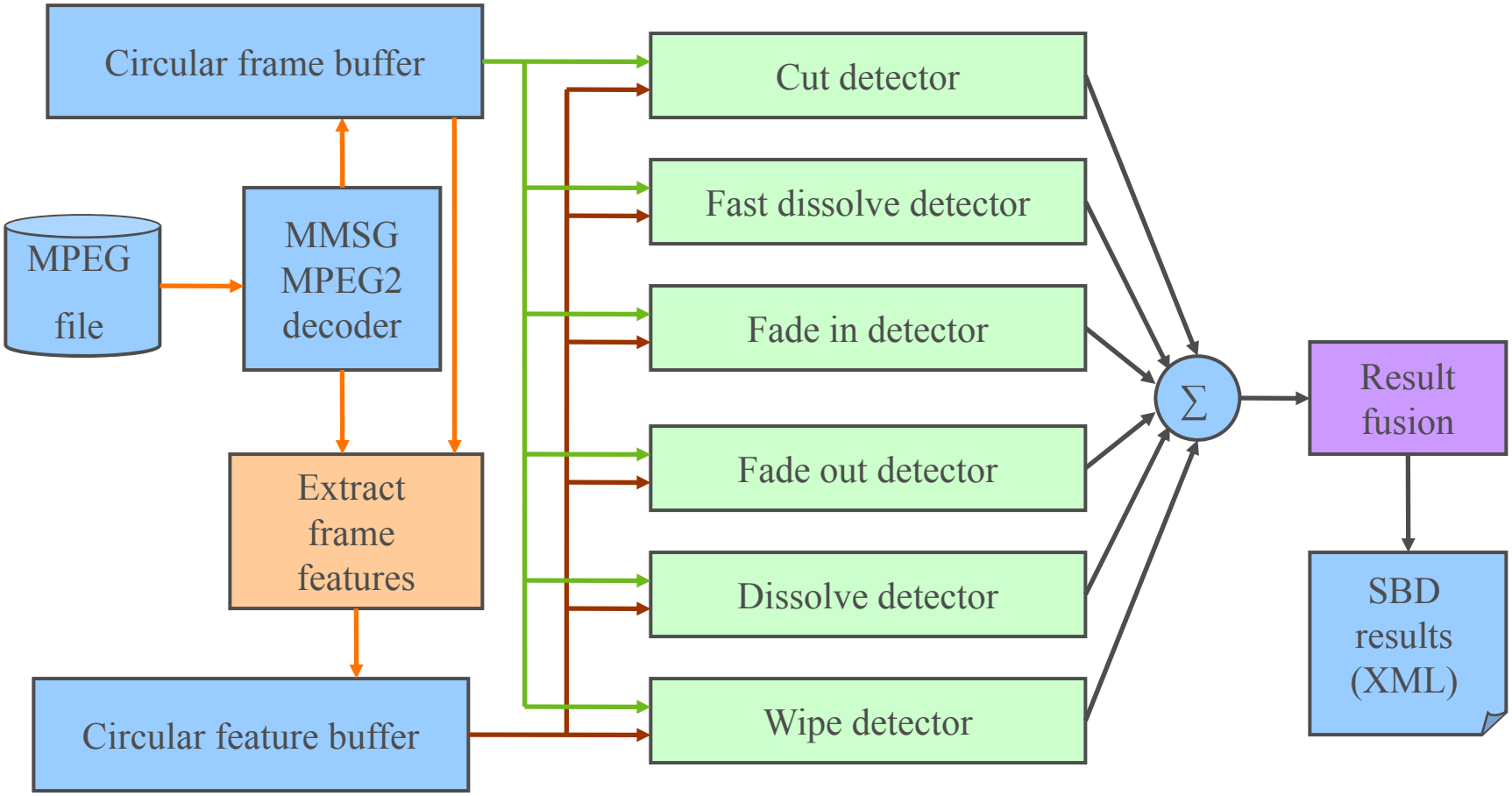
Outline

- Introduction
- **Our approach**
  - Shot boundary segmentation
  - Clustering with priors
  - Importance scoring
  - Intuitive rendering
- Evaluation results
- Conclusion

Overall Diagram of Our Approach

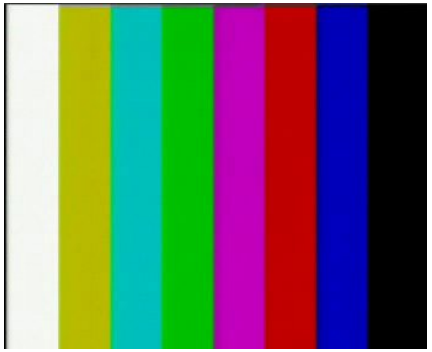


Segmentation: Shot Boundary Detection



Clustering: Junk Frame Removal

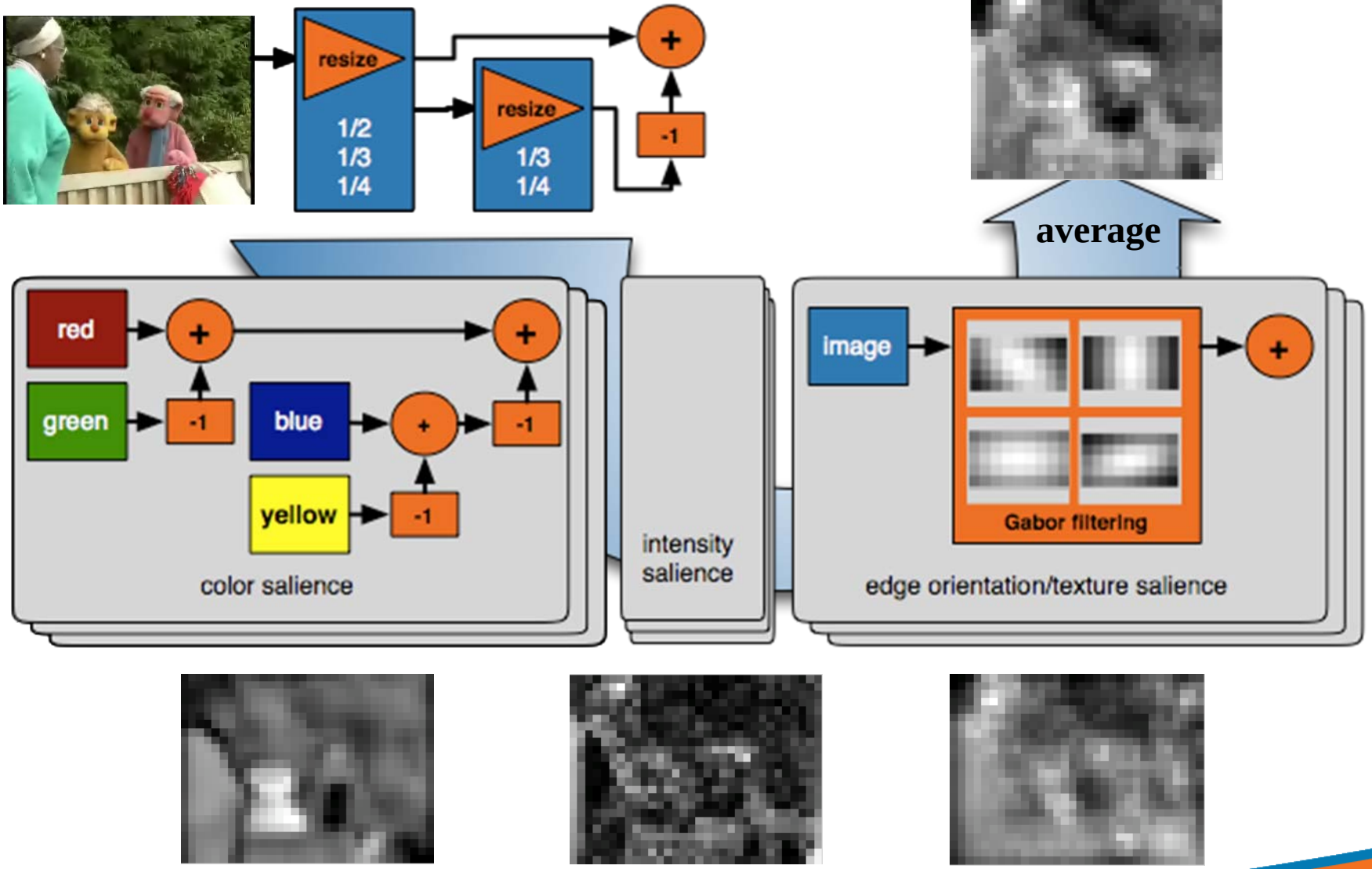
- Three main types of junk frames
  - TV test signals (color bars)
    - Audio: detect the tone signal
    - Visual: edge direction, color histogram
  - Monochrome frame
    - Visual: color histogram, intensity variance
  - Clapper frame
    - Audio: clapper sound pattern
    - Intentionally used basic visual similarity (color features) for clapper detection



Clustering:
Take Discovery from Subshots

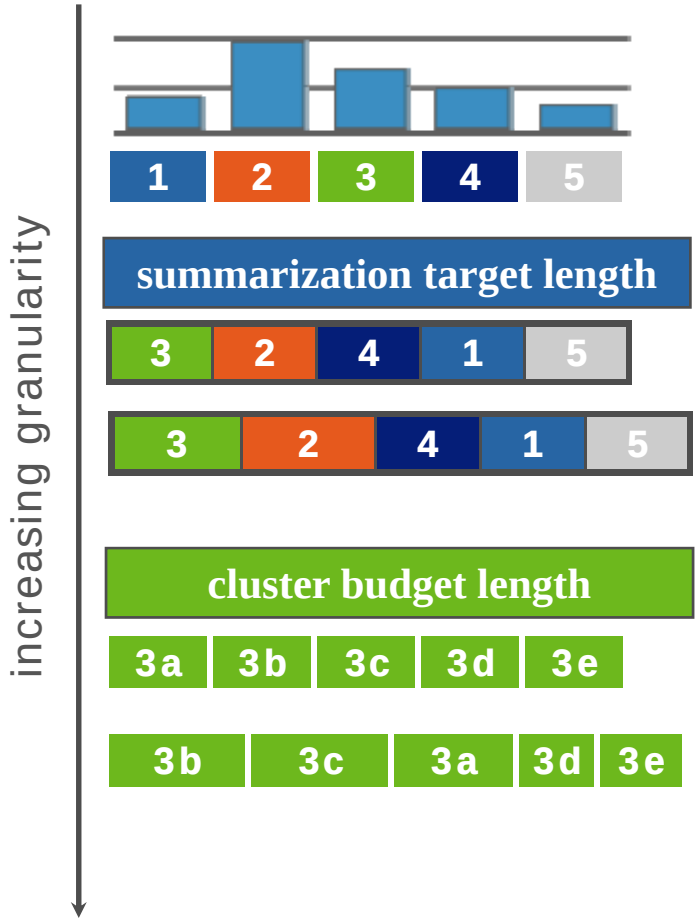
- Long shots are segmented into up to 3 subshots (uniformly sampled)
- Clustering features: grid color moment
- K-mean with two step post processing
  - Merge: compute the distance of two clusters based on motion compensated matching / histogram difference,
    - merge if the distance is small
  - Split: check the JPEG file size of each thumbnail images, split if the file sizes are too dissimilar

Importance Determination: Saliency



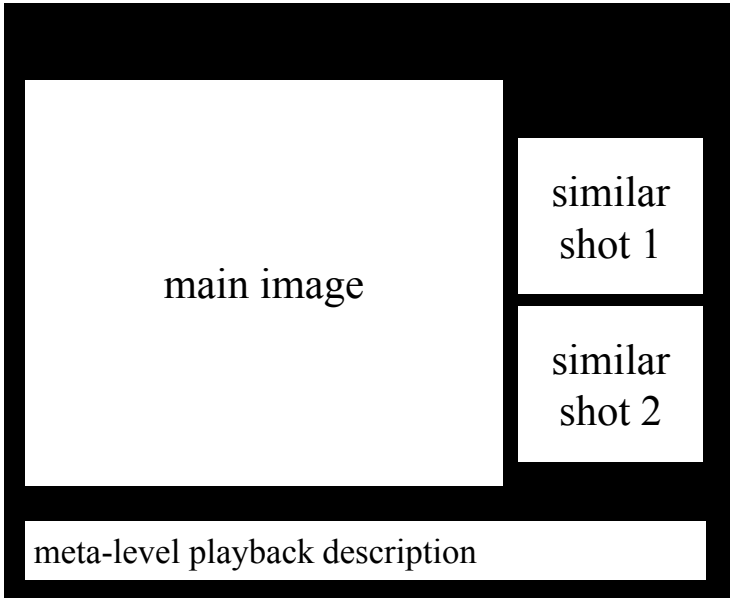
Rendering: Budget-based Allocation

- Cluster budgeting
  - Sort by importance
  - Splice time among all clusters (duration / clusters)
  - Assign remaining time by one second intervals to each cluster
- Budget subshots within clusters
  - Divide cluster budget among subshots
  - Assign additional time by importance ranked order of subshots
- Select best video segment
  - Ignore first 3 seconds if 6 second shot
  - Select longest and highest ranking contiguous region



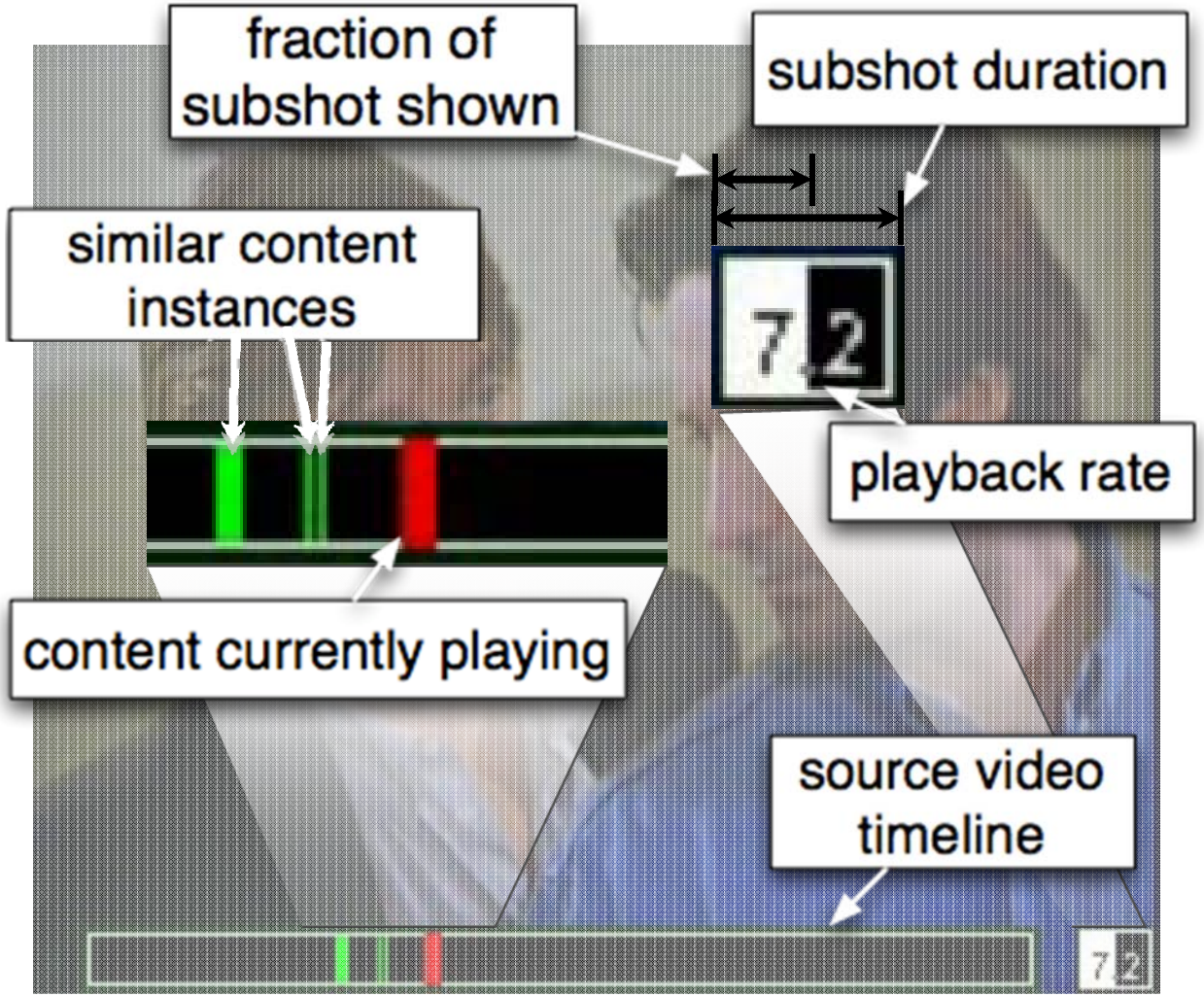
Rendering Lessons: What's not intuitive!?

- Prior approach: montage of unique shots within a cluster
  - CU submission (TRECVID 2007)
  - Confused users with multiple active shots **(too complicated)**
  - No guidance for importance or uniqueness of primary shot **(no human intuition for inter-frame matches)**



at&t

Rendering: Simplified Display



TRECVID 08
Rushes Summarization

MS221050
sum-fast run
length: 00:01.009

MS221050

TRECVID 08
Rushes Summarization

MS221050
sum-base run
length: 00:01.009

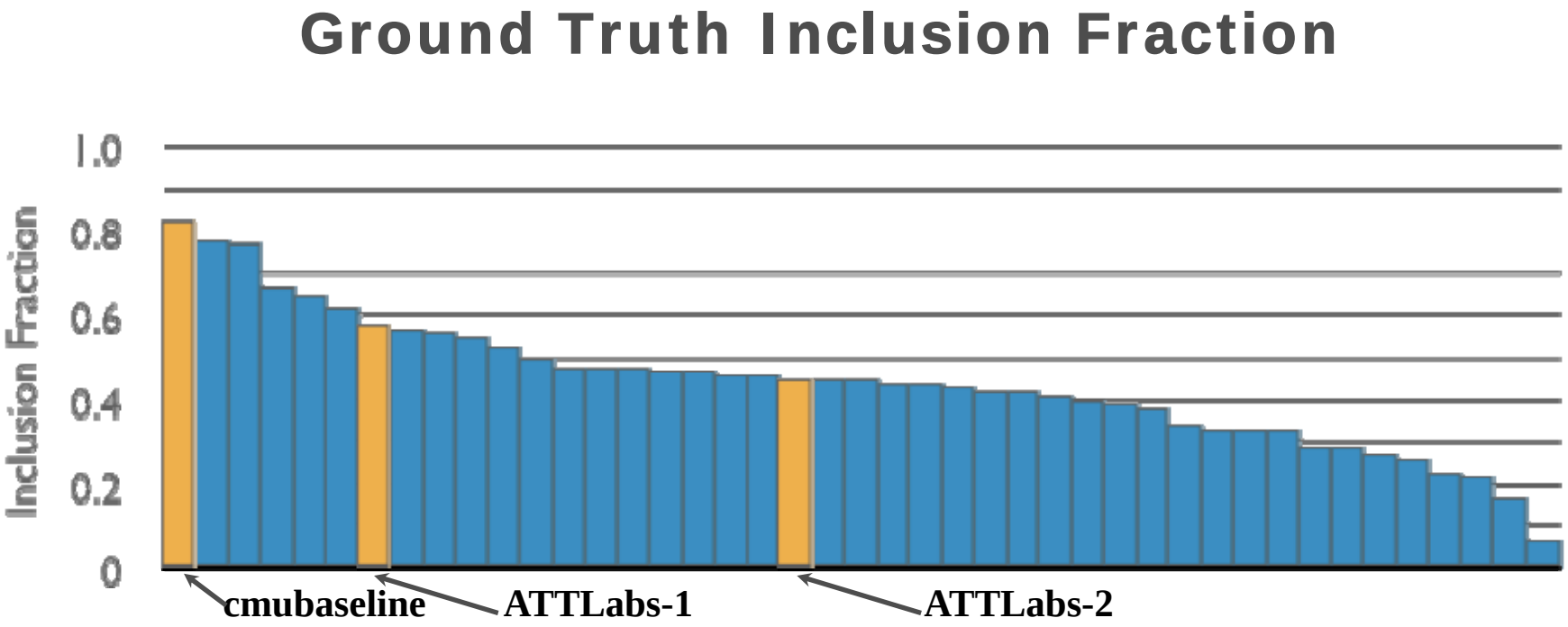
Outline

- Introduction
- Our approach
- **Evaluation results**
  - Submission descriptions
  - Analyzing assessments
- Conclusion

Submission Definitions

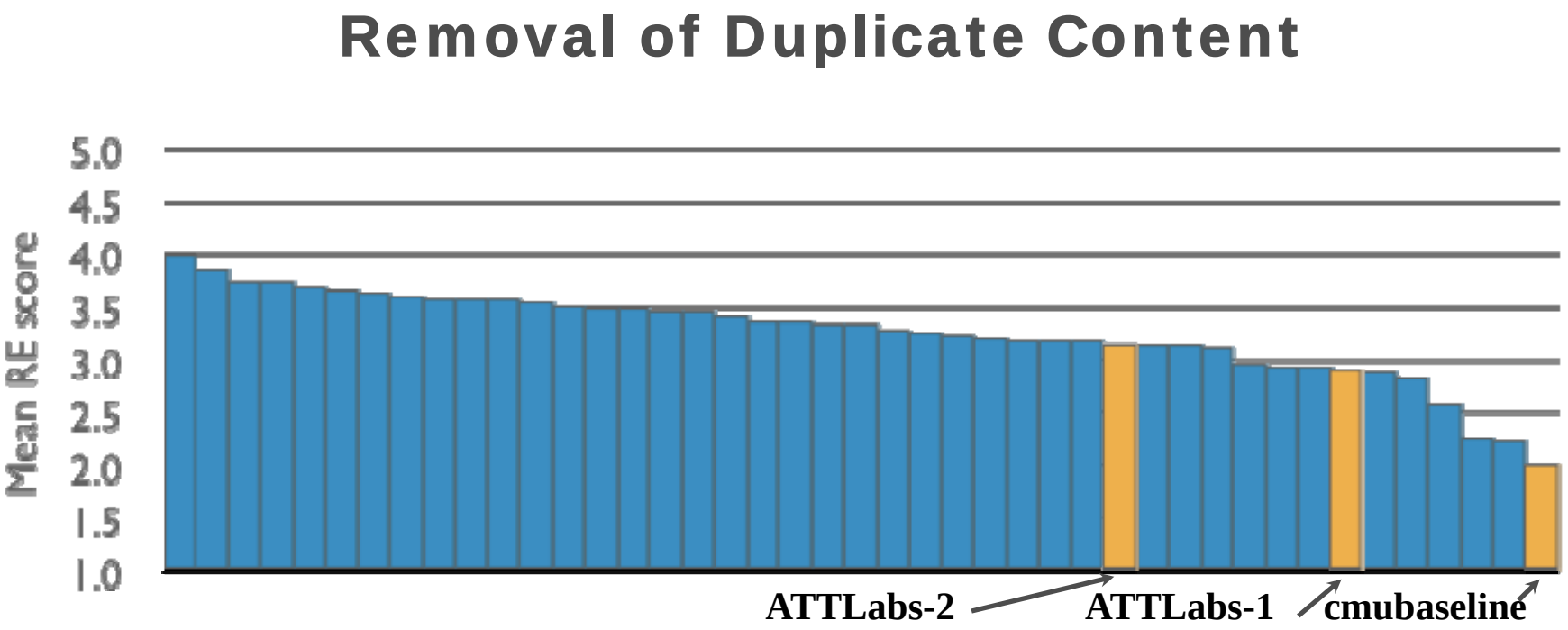
- ATTLabs-1 - dynamic speed
  - Greedy region selection constrained by shot definition, not a fixed time interval
  - Audio speed not modified; starts at clip beginning
- ATTLabs-2 - fixed speed
  - Greedy region selection in one second intervals
  - Constant (normal) speed playback
  - *Compare performance among peers to give context because of evaluation's subjective nature*
- Goal: Analyze impact of dynamic speed playback and importance (scoring) criterion

Evaluation (1)



- Increased playback rate captures more relevant events
  - Run1 and Naive CMU baseline benefit
- Base criterion for importance scoring better than median, even at constant playback rate

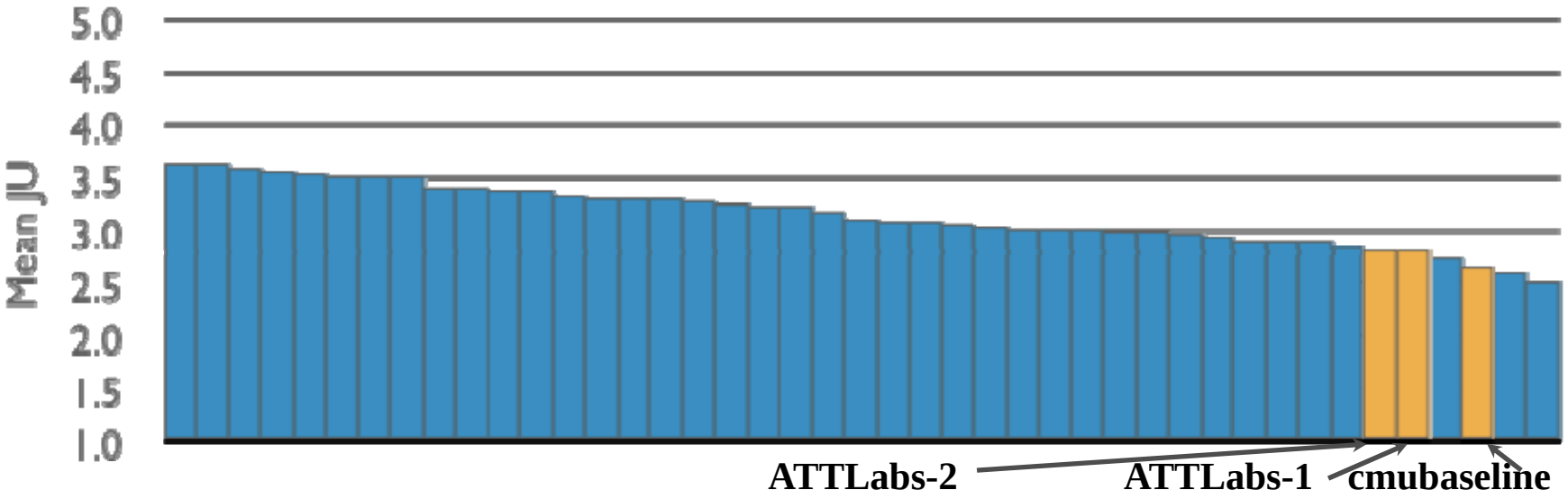
Evaluation (2)



- Clustering algorithm may be too tightly tuned to expected number of takes
- Both methods “neutral” to user with score approaching 3
  - Significantly better than baseline, which had strong belief that duplicate content was too frequent

Evaluation (3)

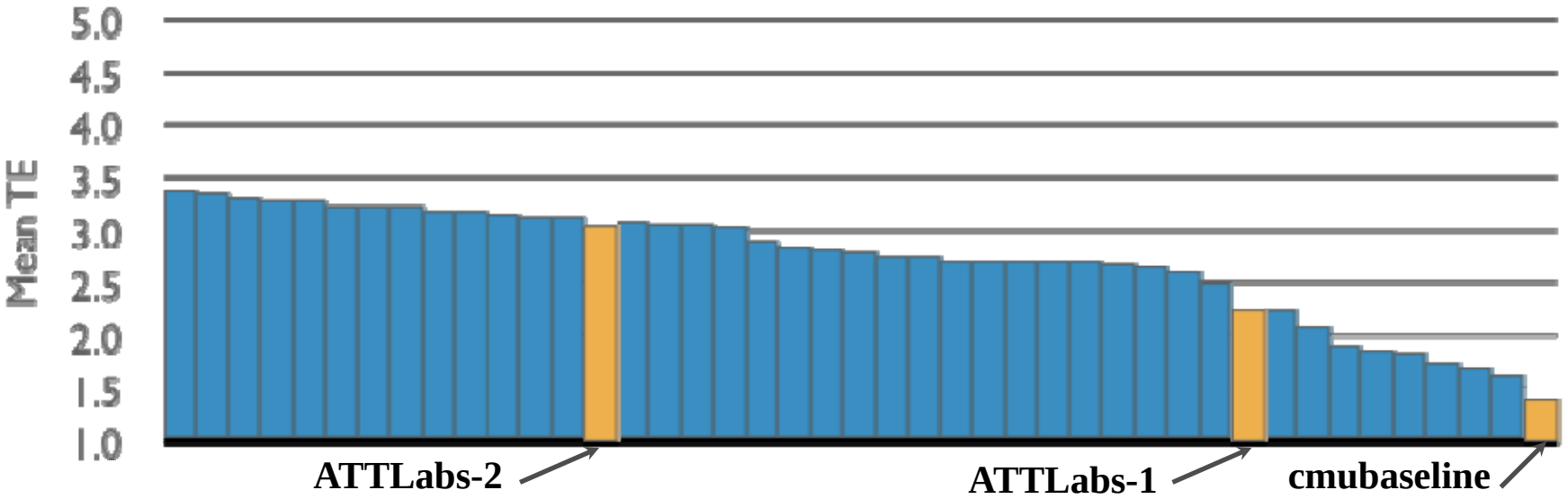
Removal of Junk (non-relevant) Content



- Efficiently removed color-bar and solid-color frames
- Clapper detection based on color and audio was not sufficient - apparent in both variable and constant speeds
- High amount of junk did not adversely effect our runs with high inclusion rate

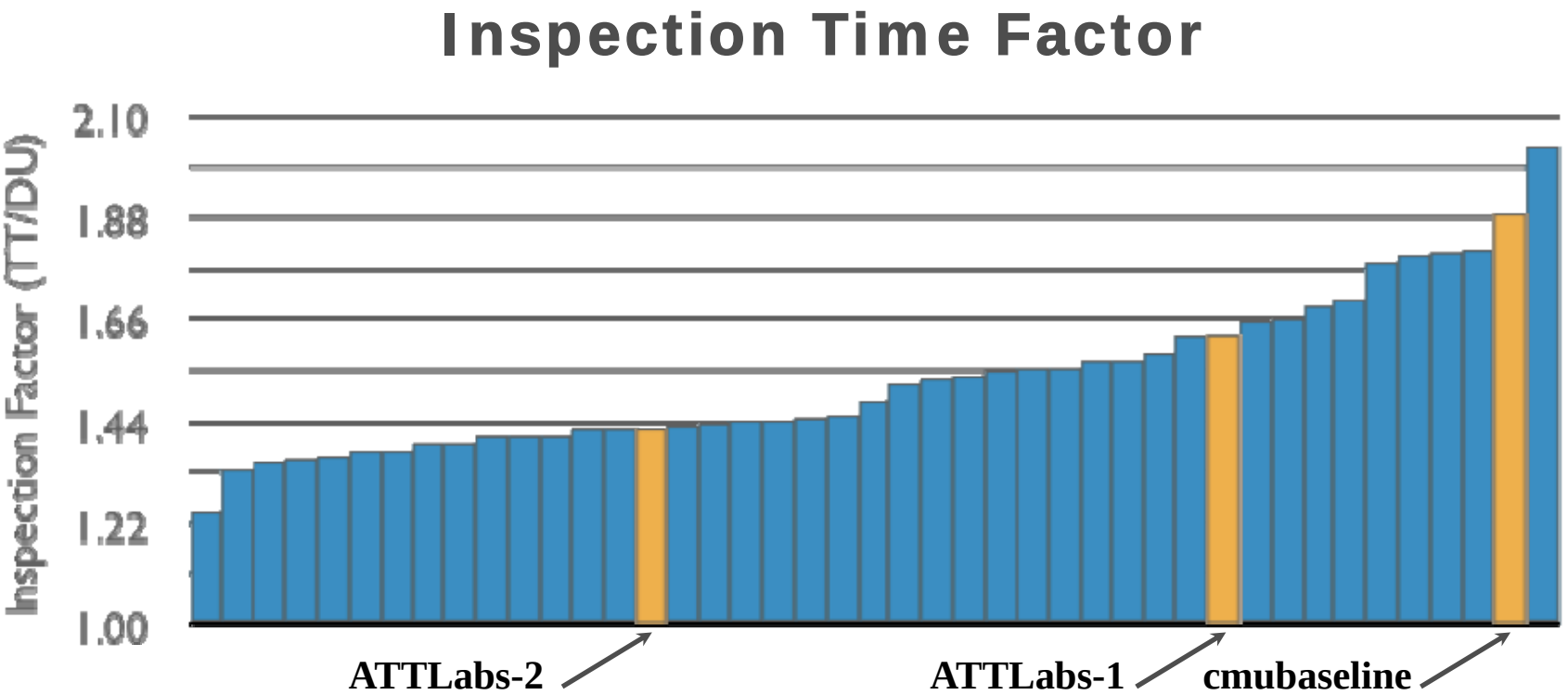
Evaluation (4)

Preference for Visual Tempo



- No strong positive opinions evoked by any system
- Constant-rate playback was generally acceptable
  - Implicit acceptance of rendering display
- Non-constant speed systems had jarring/disliked tempo
  - Variable rate better than fixed high-speed

Evaluation (5)



- Increased playback rate requires more time for inspection
- **Factor** for constant speed less than median factor
  - Reasonable for inspection time to be longer than duration because of high truth inclusion (see eval.1)

Conclusions

- New model for user-centric summaries fares well compared to data-centric models
- Shot-segmentation and budgeting method accurately picks important regions
- Users can tolerate and understand high-speed video playback
- Importance scores approximated well by emulating human interest and motion
- Joint A/V techniques are helpful on some junk material, but more intricate scene/object method needed for clappers

Conclusion: Future Work

- Clustering
  - Enhance A/V clustering with visual SIFT for identification of diverse junk-frame content
- Importance scoring
  - Analyze changes in salience over time
  - Visually assisting users in salient object or event identification
- Rendering
  - Apply limits on speed up allowed in time-dynamic rendering
  - Analyze impact of intelligent audio selection using acoustic salience (i.e. intensity/prosody)

Assisting users in interesting object identification
(a la modern crime drama)