Extreme Video Retrieval

Maximizing the Synergy between Systems and Humans

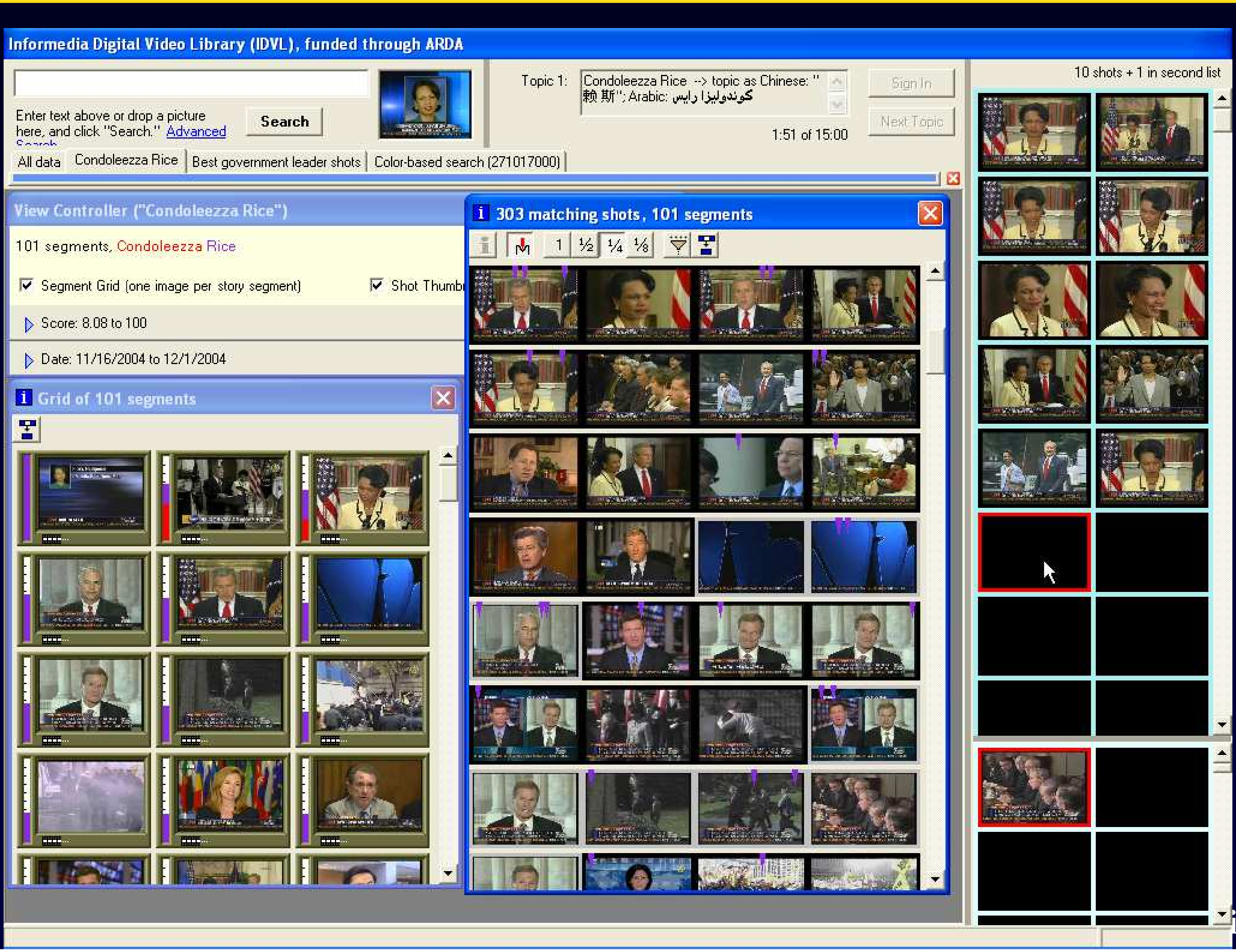
TRECVID meeting – November 15, 2005

The Informedia Team
Carnegie Mellon University
Pittsburgh, USA

“Classic Informedia” Interface Work

- Interactive Video Queries
  - Multilingual and fielded text query matching capabilities
  - Faster color-based matching with simplified interface for launching color queries
- Interactive Browsing, Filtering, and Summarizing
  - Browsing by person-in-the-news
  - Browsing by visual concepts
  - Quick display of contents and context in synchronized views
- Testing with Novice Users as well as Experts
  - Same questionnaires used as with TRECVID 2004 (to get satisfaction usability measure and help interpret results)
  - Logging to test “Extreme Light” interface supporting text, color, and concept browsing/search

TRECVID Evaluation Interface Example



Informedia Digital Video Library (IDVL), funded through ARDA

Enter text above or drop a picture here, and click "Search." Advanced Search

Topic 1: Condoleezza Rice --> topic as Chinese: " 賴 斯"; Arabic: كوندوليزا رايس

1:51 of 15:00

10 shots + 1 in second list

View Controller ("Condoleezza Rice")

101 segments, Condoleezza Rice

☑ Segment Grid (one image per story segment)

☑ Shot Thumb

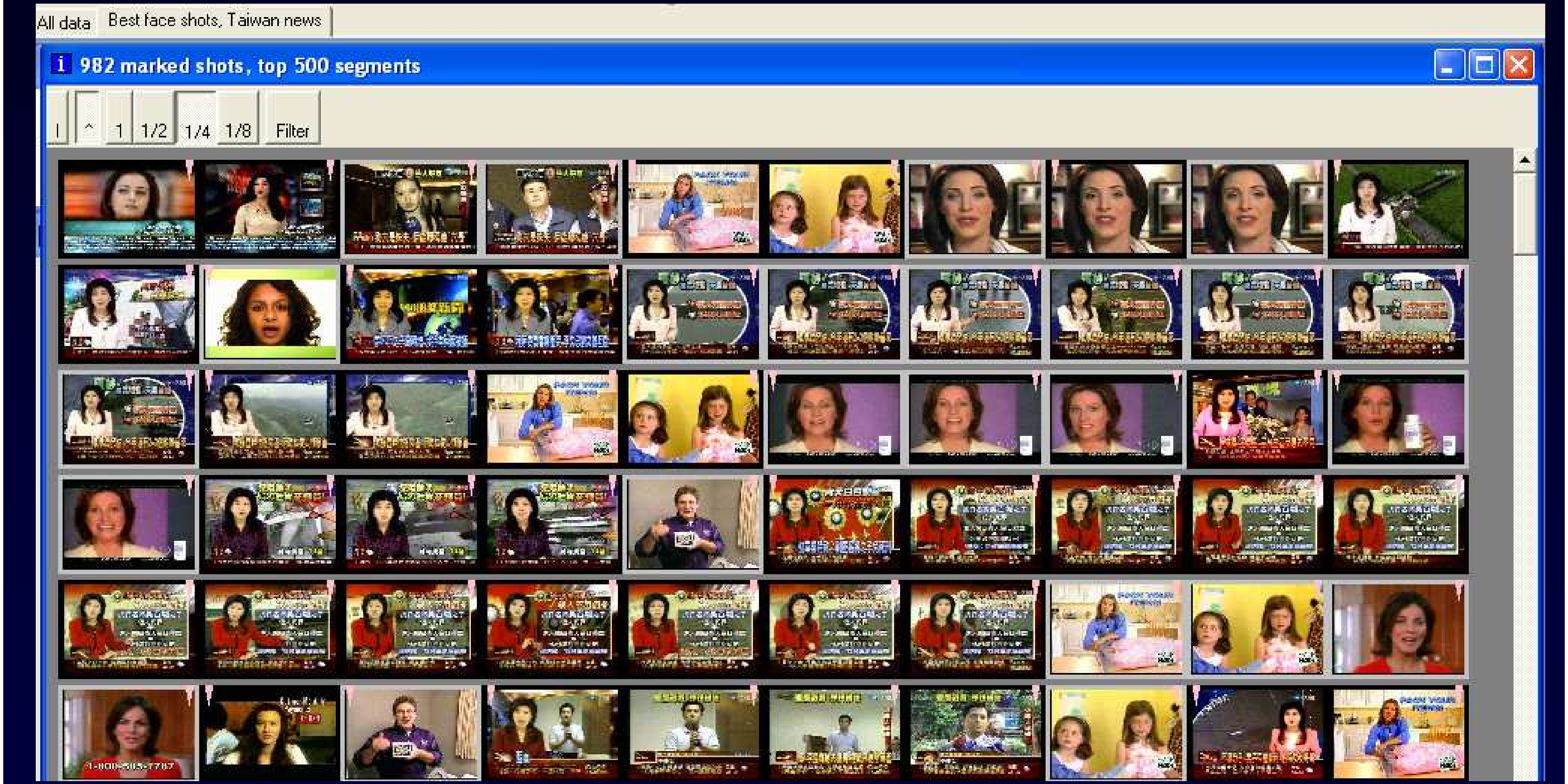
Score: 8.08 to 100

Date: 11/16/2004 to 12/1/2004

Grid of 101 segments

303 matching shots, 101 segments

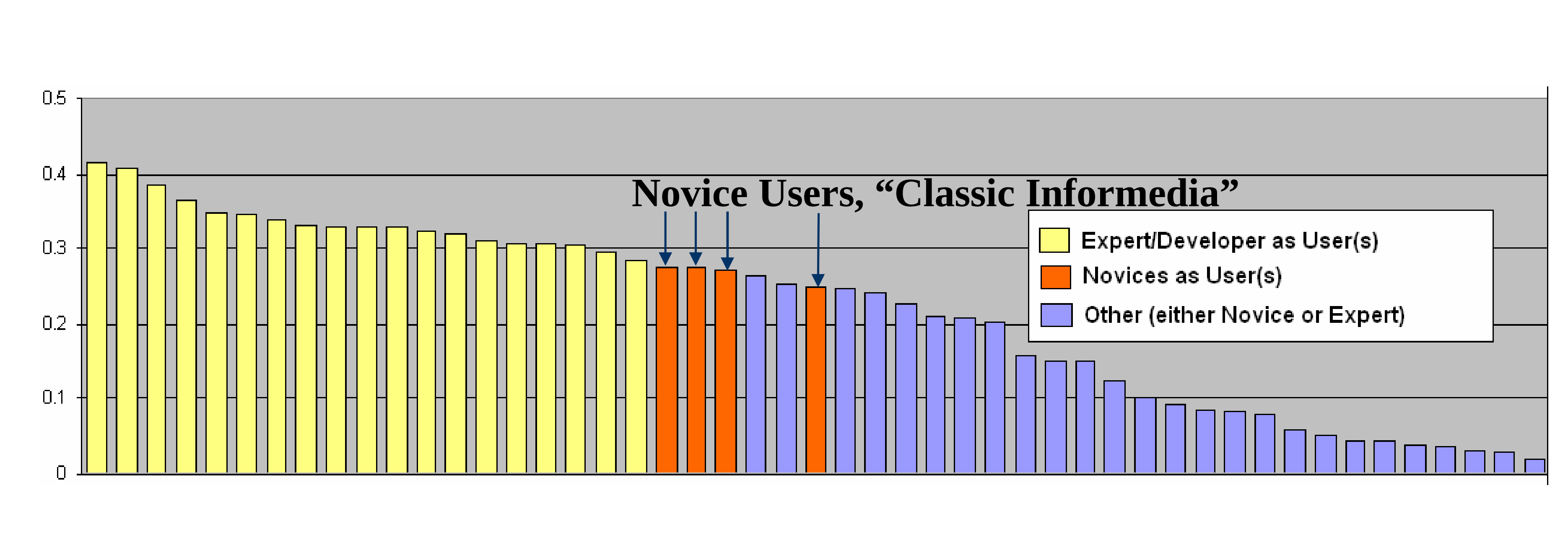
Visual Browsing


All data
Best face shots, Taiwan news
982 marked shots, top 500 segments
|
^
1
1/2
1/4
1/8
Filter

“Classic Informedia” Results

- Concept browsing and image search used much more relative to text search compared to prior TRECVIDs
- Novices still have lower performance than experts (reconfirming 2004 studies, with logs of actions for follow-up analysis)
- Nature of topics caused “interactive” this year to be more one-shot query, less browsing/exploration
    - Performance improvements not found for leveraging usage context (hiding shots judged in prior queries)
    - “Extreme-light” interface including concept browsing often good enough that user never proceeded on to any query
- “Classic Informedia” scored highest of those testing with novice users

TRECVID’05 Interactive Search Results

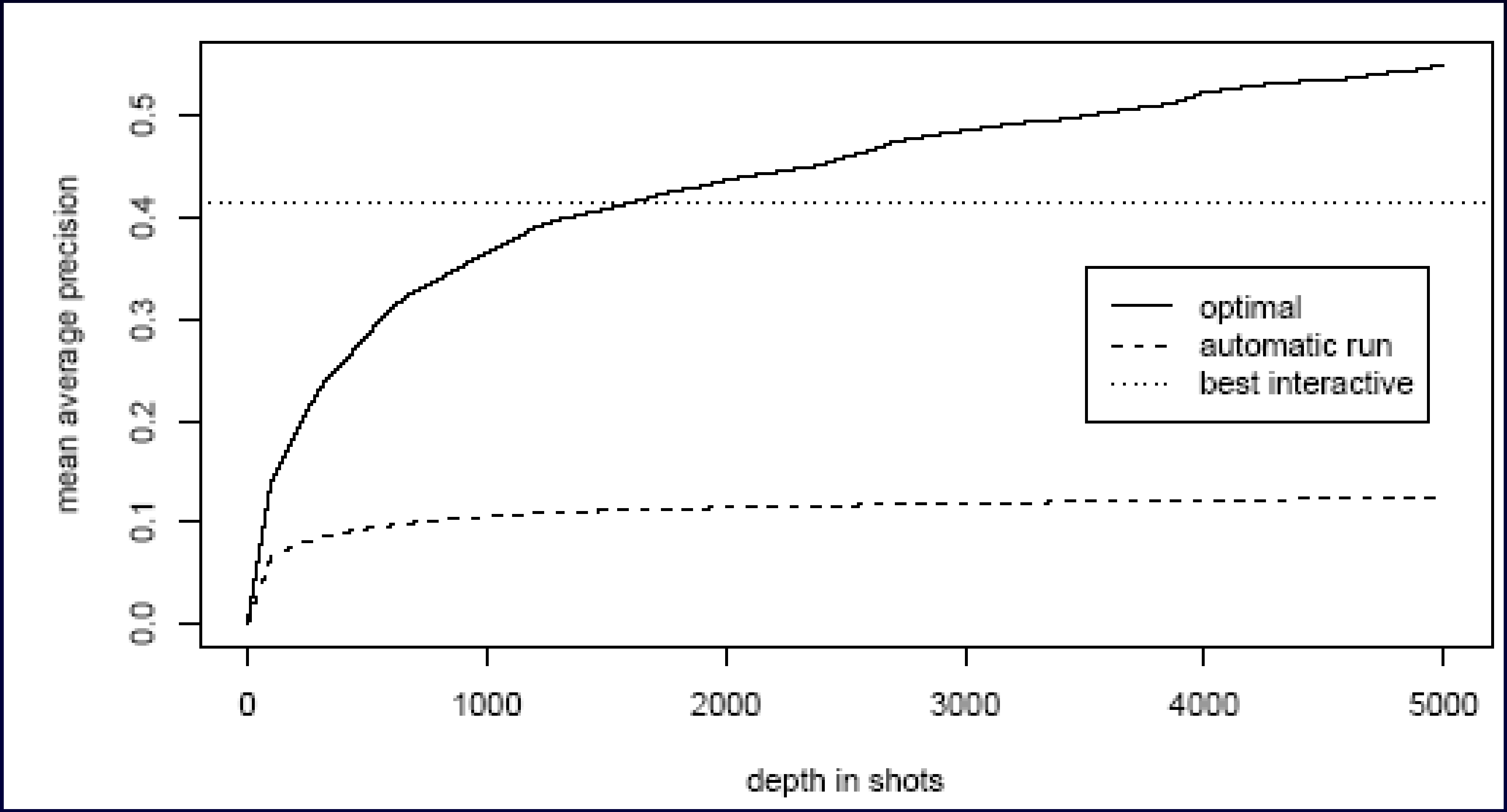


The Goal of Extreme Video Retrieval

Exploring Video Search at the Limits of Human and System Performance

A Different Approach

Observations about Automatic vs Interactive Search



Extreme Video Retrieval

- Automatic retrieval baseline for ranked shot order

- Two methods of presentation:

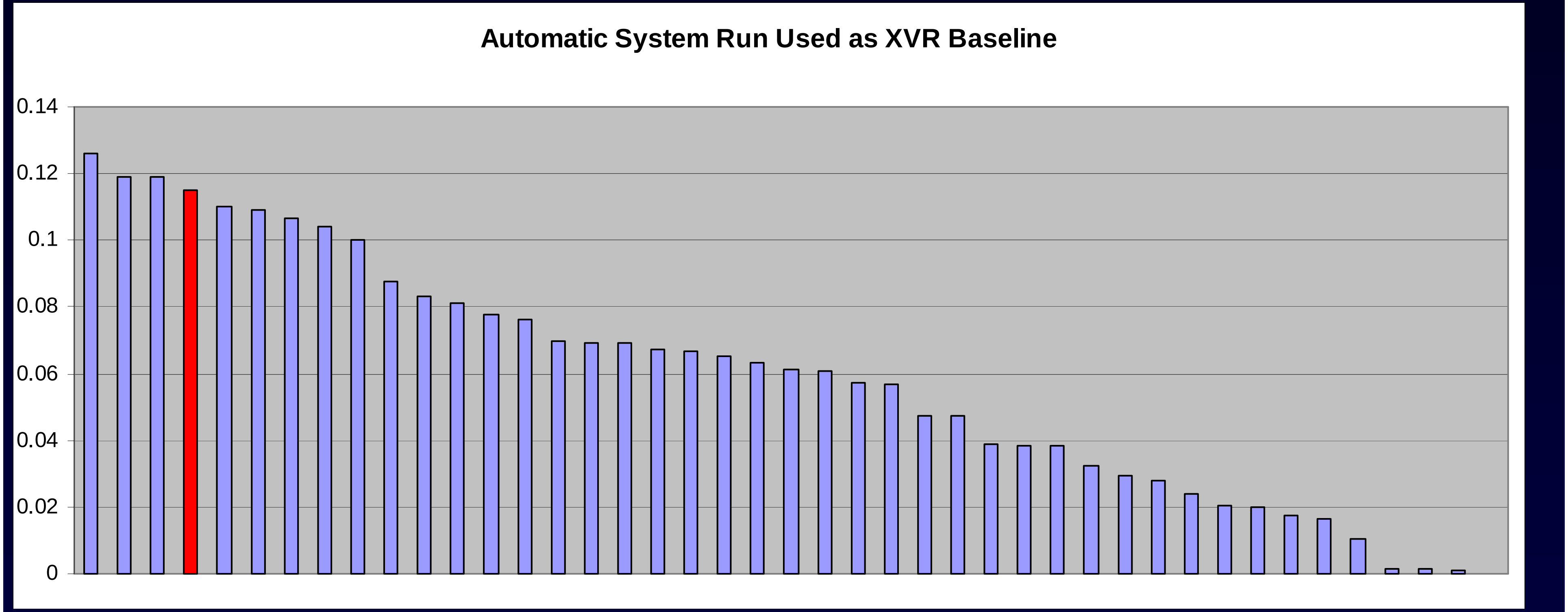
  User-controlled or *System-controlled* time interval

  - User-controlled Presentation – Manual Browsing with Resizing of Pages
  - System-controlled Presentation - Rapid Serial Visual Presentation (RSVP)

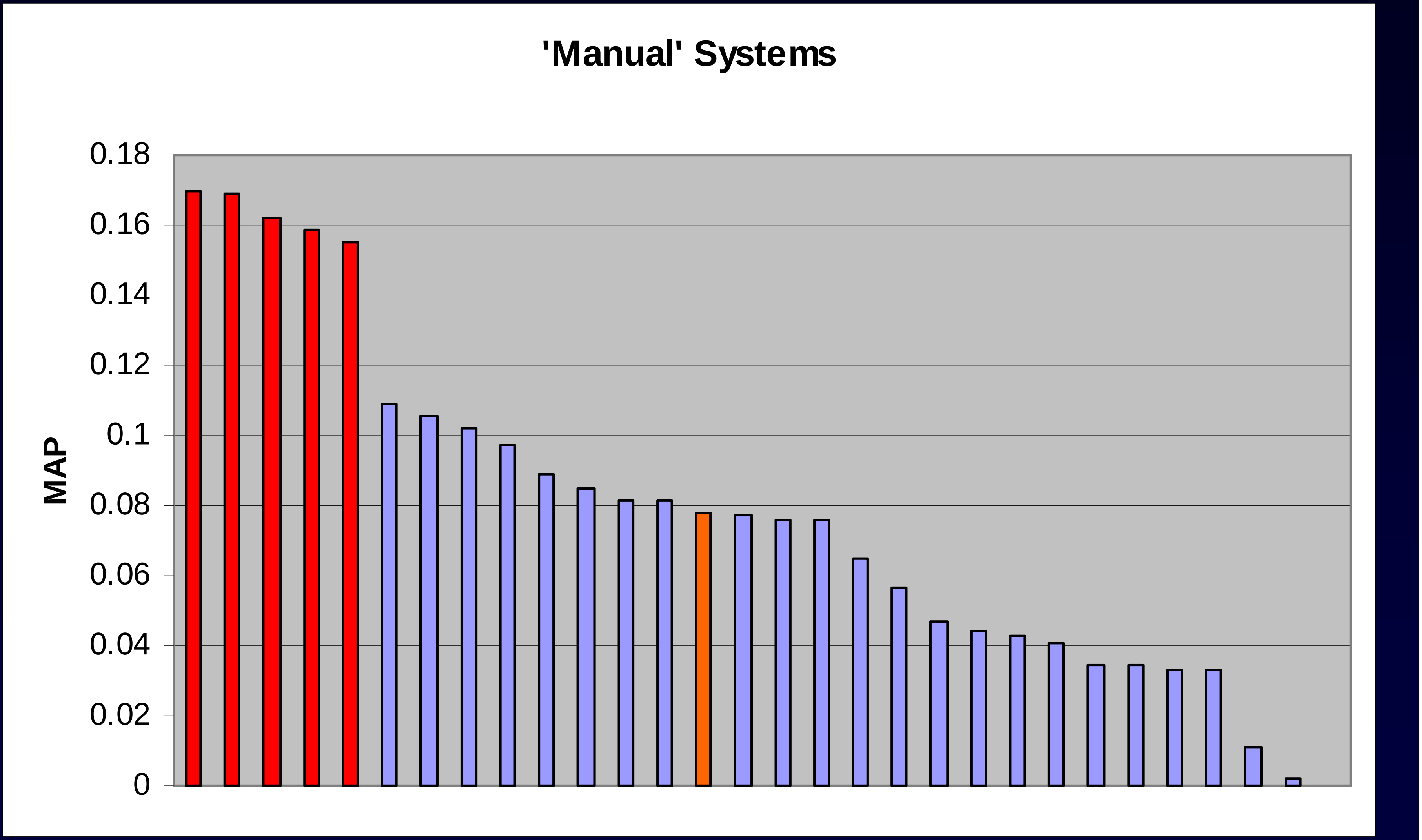
The Automatic System Result

- Start with automatic system generated result
- 5 uni-modal retrieval “experts” and 15 semantic features
  - Experts: Text, Color, Texture, Edge, PersonX
  - Features: Face, Anchor, Commercial, Studio, Graphics, Weather, Sports, Outdoor, Person, Crowd, Road, Car, Building, Motion
- A relevance-based probabilistic retrieval model
  - Basic model: “ranking” logistic regression
    - Reduce the disorders between positive/negative data
  - Query analysis: incorporate the query information into the combination function
    - Five query types with combination weights learned from TREC04
- Present shots (image keyframes) in ranked order

XVR Automatic Baseline (Unevaluated)



TRECVID Manual Results



User-controlled presentations

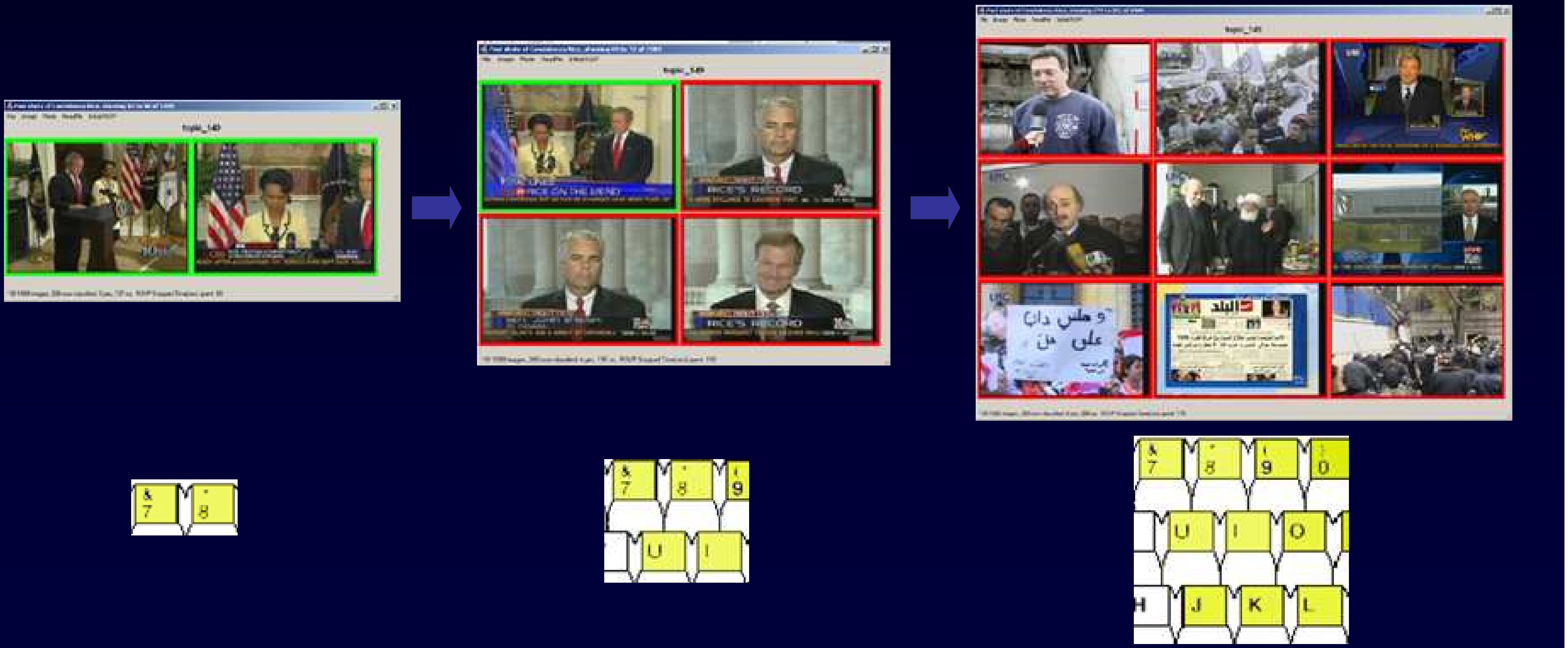
- Manual Browsing with Resizing of Pages
  - Manually page through images
    - User decides to view next page
  - Vary the number of images on a page (2, 4, 9, 16)

  - Allow chording on the keypad to identify shots of interest

- Also tried clustering by story and without resizing of pages
  - Not as effective

- A very brief final verification step (1 min)

MBRP - Manual Browsing with Resizable Pages

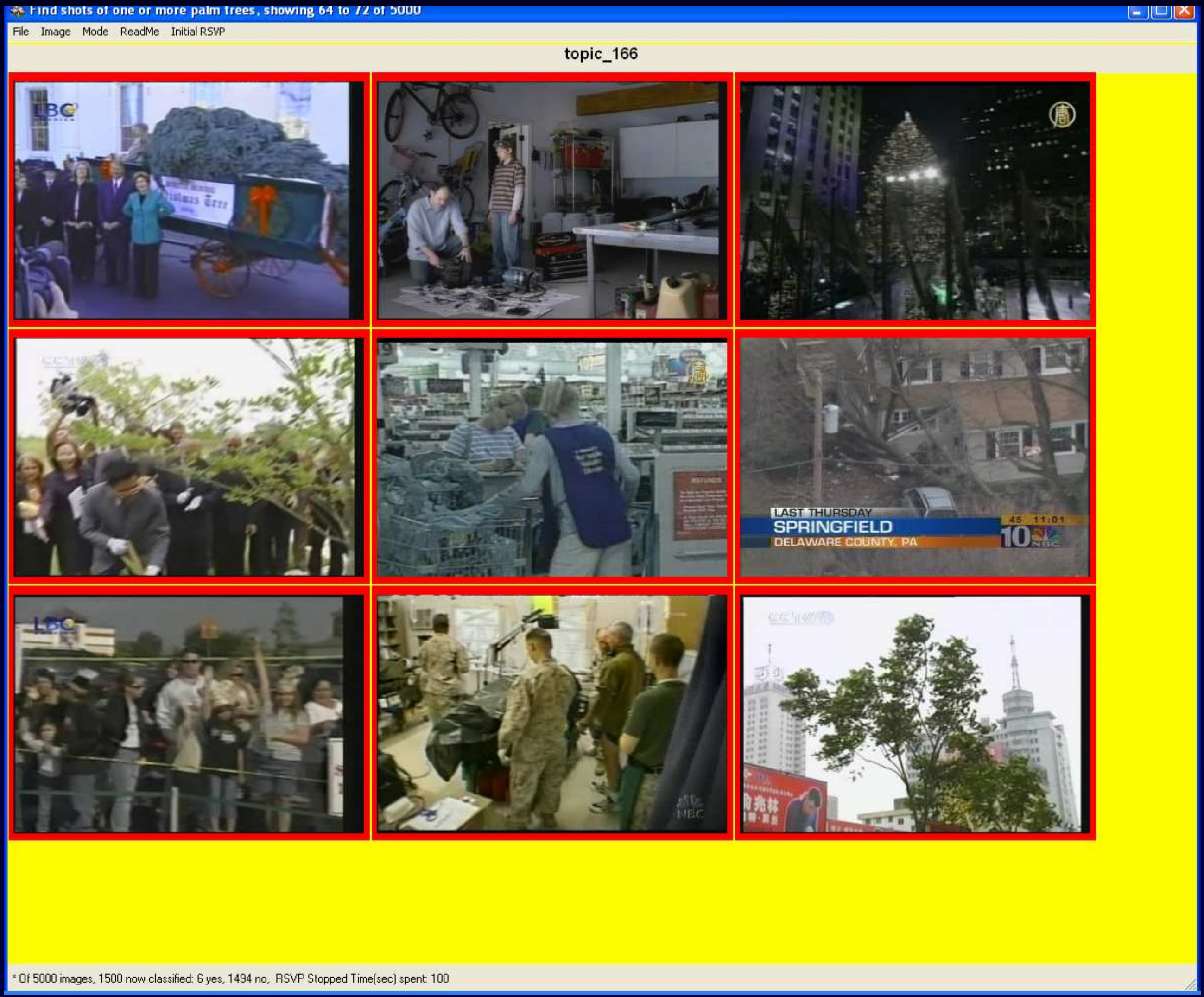


System-controlled Presentation

- Rapid Serial Visual Presentation (RSVP)
  - Minimizes eye movements
    - All images in same location
  - Maximizes information transfer: System à Human
    - Up to 10 key images/second
    - 1 or 2 images per page
    - Presentation intervals are dynamically adjustable by the user
      - Slower initially (or when “breaks” are needed)
        - Many relevant images, user needs habituation
      - Faster after a few minutes (100 msec/page increments)
        - Few relevant images, accommodation
  - Click when relevant shot is seen
    - Mark previous page also as relevant

- A final verification step (~3 min) is necessary
  - Should be related to the number of relevant shots

Extreme QA with RSVP



Find shots of one or more palm trees, showing 64 to 72 of 5000

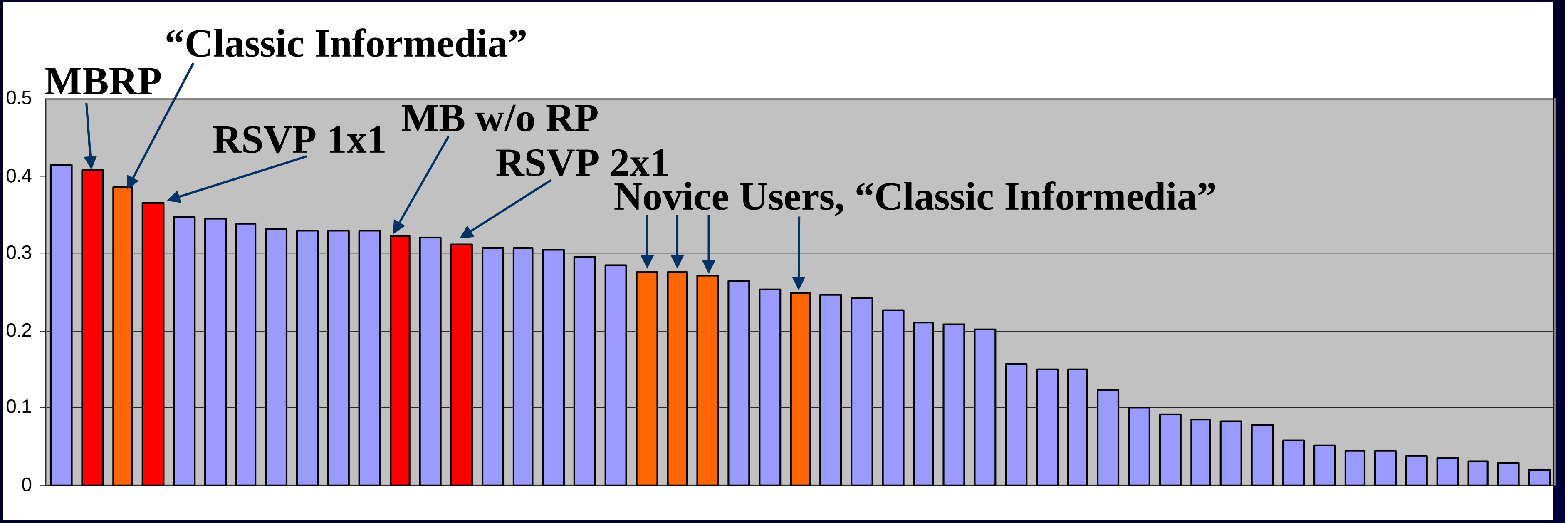
* Of 5000 images, 1500 now classified: 6 yes, 1494 no, RSVP Stopped Time(sec) spent: 100

3x3 display

1 page/second

Numpad chording to select shots

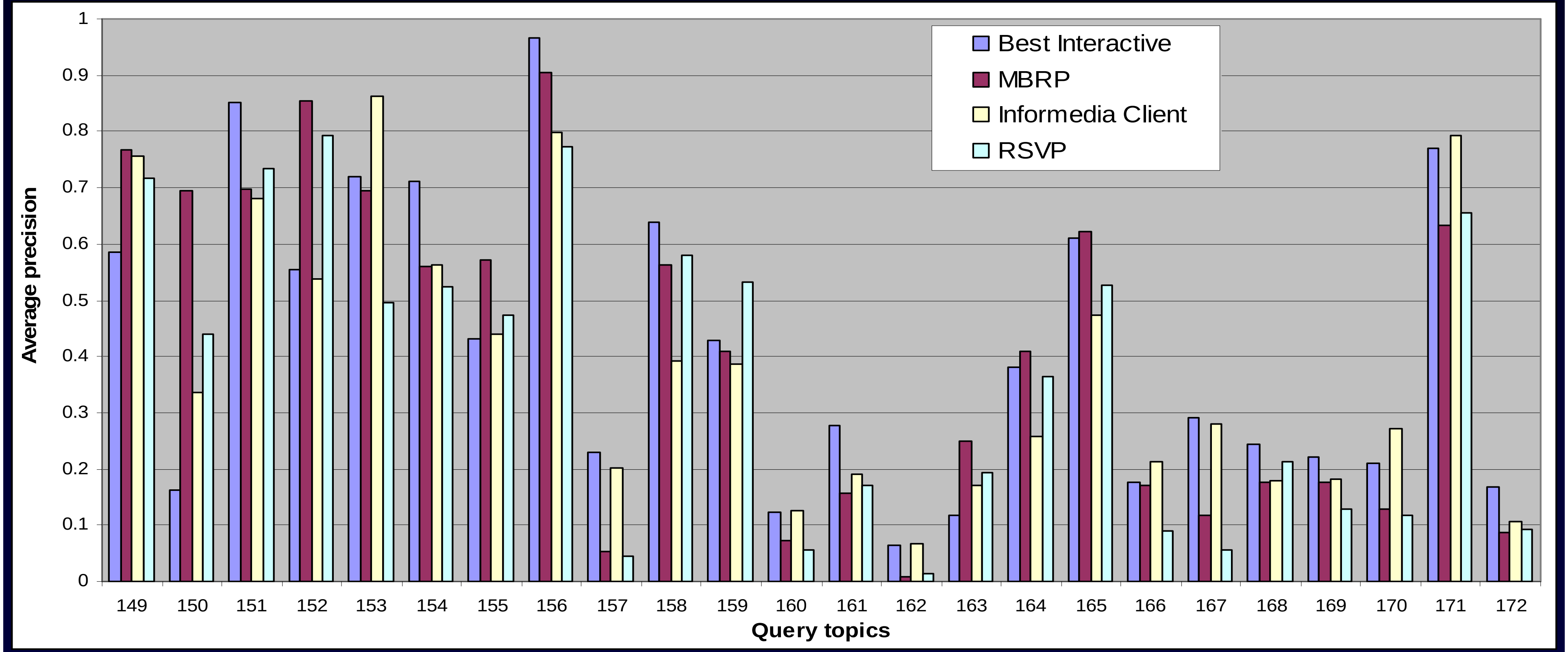
Informedia TRECVID'05 Interactive Search Results



ARDA

Carnegie Mellon

TRECVID’05 Interactive Results by Topic



The Future of Extreme Video Retrieval

Eventually, we envision the computer will observe the user and LEARN!

The system can learn:

- What object and image characteristics are relevant
- What text characteristics (words) are relevant to the query
- What combination weights should be used to combine them

Based on shots that have just been marked as relevant

- As learning improves, the human has to do less and less work

We exploit the human’s ability to quickly mark relevant shots and the computer’s ability to learn from given examples

Questions?

Thank You

Carnegie Mellon University