Carnegie Mellon University
Search

TRECVID 2004 Workshop – November 2004

Mike Christel, Jun Yang, Rong Yan, and Alex Hauptmann
Carnegie Mellon University
christel@cs.cmu.edu
Talk Outline

• CMU Informedia interactive search system features
• 2004 work: novice vs. expert, visual-only (no audio processing, hence no automatic speech recognized [ASR] text, no closed-captioned text) vs. full system that does use ASR and CC text
• Examination of results, esp. of visual-only vs. full system
  • Questionnaires
  • Transaction logs
• Automatic and manual search
• Conclusions
Informedia Acknowledgments

• Supported by the Advanced Research and Development Activity (ARDA) under contract number NBCHC040037 and H98230-04-C-0406

• Contributions from many researchers – see http://www.informedia.cs.cmu.edu for more details
CMU Interactive Search, TRECVID 2004

• Challenge from TRECVID 2003: how usable is system without the benefit of ASR or CC (closed caption) text?
  • Focus in 2004 on “visual-only” vs. “full system”
  • Maintain some runs for historical comparisons
• Six interactive search runs submitted
  • Expert with full system (addressing all 24 topics)
  • Experts with visual only system (6 experts, 4 topics each)
  • Novices, within-subjects design where each novice sees 2 topics in “full system” and 2 in “visual-only”
    - 24 novice users (mostly CMU students) participated
    - Produced 2 “visual-only” runs and 2 “full system” runs
Two Clarifications

• Type A or Type B or Type C?
  • Marked search runs as Type C **ONLY** because of the use of a face classifier by Henry Schneideman which was trained with non-TRECVID data
  • That face classification provided to TRECVID community
• Meaning of “expert” in our user studies
  • “Expert” meant expertise with the Informedia retrieval system, **NOT** expertise with the TRECVID search test corpus
  • “Novice” meant that user had no prior experience with video search as exhibited by the Informedia retrieval system nor any experience with Informedia in any role
• **ALL** users (novice and expert) had no prior exposure to the search test corpus before the practice run for the opening topic (limited to 30 minutes or less) was conducted
Interface Support for Visual Browsing
Interface Support for Image Query

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

All data  Image Search [413312]
Interface Support for Text Query

192 results for "fire explosion bombing"

384 matching shots, 192 segments
Interface Support to Filter Rich Visual Sets

All data  Best building shot:

1 562 marked shots, 344 segments

Filter  Collapse Mode

Shots shown:

76 / 562

no faces  no people  no building

only faces  only people  only building

only anchors  only ads  only outdoor

ARDA  Carnegie Mellon
Characteristics of Empirical Study

- 24 novice users recruited via electronic bboard postings
- Independent work on 4 TRECVID topics, 15 minutes each
- Two treatments: F – full system, V – visual-only (no closed captioning or automatic speech recognized text)
- Each user saw 2 topics in treatment “F”, 2 in treatment “V”
- 24 topics for TRECVID 2003, so this study produced four complete runs through the 24 topics: two in “F”, two in “V”
- Intel Pentium 4 machine, 1600 x 1200 21-inch color monitor
- Performance results remarkably close for the repeated runs:
  - 0.245 mean average precision (MAP) for first run through treatment “F”, 0.249 MAP for second run through “F”
  - 0.099 MAP for first run through treatment “V”, 0.103 MAP for second run through “V”
A Priori Hope for Visual-Only Benefits

Optimistically, hoped that visual-only system would produce better avg. precision on some “visual” topics than full system, as visual-only system would promote “visual” strategies.
Novice Users’ Performance

TRECVID 2004 Interactive Search, Novices with either Full or Visual-Only System

- Full1
- Full2
- VisualOnly1
- VisualOnly2

Average Precision

125 street 126 flood 127 dog 128 Hyde 129 Dome 130 hockey 131 keys 132 stretcher 133 Saddam 134 Yeltsin 135 Donaldson 136 golf 137 Netanyahu 138 steps 139 weapon 140 bikes 141 umbrellas 142 tennis 143 wheelchairs 144 Clinton 145 horses 147 fire 148 signs
Expert Users’ Performance

TRECVID 2004 Interactive Search, Expert with either Full or Visual-Only System

- **Expert, Full**
- **Expert, Visual-Only**

**Average Precision**

125 street, 126 flood, 127 dog, 128 Hyde, 129 Dome, 130 hockey, 131 keys, 132 stretcher, 133 Saddam, 134 Yeltsin, 135 Donaldson, 136 golf, 137 Netanyahu, 138 steps, 139 weapon, 140 bikes, 141 umbrellas, 142 tennis, 143 wheelchairs, 144 Clinton, 145 horses, 147 fire, 148 signs
Mean Avg. Precision, TRECVID 2004 Search

137 runs (62 interactive, 52 manual, 23 automatic)
TRECVID04 Search, CMU Interactive Runs

CMU Expert, Full System
CMU Novice, Full System
CMU Expert, Visual-Only
CMU Novice, Visual-Only

Interactive
Manual
Automatic
TRECVID04 Search, CMU Search Runs

CMU Expert, Full System
CMU Novice, Full System
CMU Expert, Visual-Only
CMU Novice, Visual-Only
CMU Manual
CMU Automatic

Interactive
Manual
Automatic
Satisfaction, Full System vs. Visual-Only

12 users asked which system treatment better:
- 4 liked first system better, 4 second system, 4 no preference
- 7 liked full system better, 1 liked the visual-only system better
## Summary Statistics, User Interaction Logs

<table>
<thead>
<tr>
<th>(statistics reported as averages)</th>
<th>Novice Full</th>
<th>Novice Visual</th>
<th>Expert Full</th>
<th>Expert Visual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of minutes spent per topic (fixed by study)</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Text queries issued per topic</td>
<td>9.04</td>
<td>14.33</td>
<td>4.33</td>
<td>5.21</td>
</tr>
<tr>
<td>Word count per text query</td>
<td>1.51</td>
<td>1.55</td>
<td>1.54</td>
<td>1.30</td>
</tr>
<tr>
<td>Number of video story segments returned by each text query</td>
<td>105.29</td>
<td>15.65</td>
<td>79.40</td>
<td>20.14</td>
</tr>
<tr>
<td>Image queries per topic</td>
<td>1.23</td>
<td>1.54</td>
<td>1.13</td>
<td>6.29</td>
</tr>
<tr>
<td>Precomputed feature sets (e.g., “roads”) browsed per topic</td>
<td>0.13</td>
<td>0.21</td>
<td>0.83</td>
<td>1.92</td>
</tr>
</tbody>
</table>
### Summary Statistics, User Interaction Logs

<table>
<thead>
<tr>
<th>(statistics reported as averages)</th>
<th>Novice Full</th>
<th>Novice Visual</th>
<th>Expert Full</th>
<th>Expert Visual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of minutes spent per topic (fixed by study)</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Text queries issued per topic</td>
<td>9.04</td>
<td>14.33</td>
<td>4.33</td>
<td>5.21</td>
</tr>
<tr>
<td>Word count per text query</td>
<td>1.51</td>
<td>1.55</td>
<td>1.54</td>
<td>1.30</td>
</tr>
<tr>
<td>Number of video story segments returned by each text query</td>
<td>105.29</td>
<td>15.65</td>
<td>79.40</td>
<td>20.14</td>
</tr>
<tr>
<td>Image queries per topic</td>
<td>1.23</td>
<td>1.54</td>
<td>1.13</td>
<td>6.29</td>
</tr>
<tr>
<td>Precomputed feature sets (e.g., “roads”) browsed per topic</td>
<td>0.13</td>
<td>0.21</td>
<td>0.83</td>
<td>1.92</td>
</tr>
</tbody>
</table>
### Summary Statistics, User Interaction Logs

<table>
<thead>
<tr>
<th>(statistics reported as averages)</th>
<th>Novice Full</th>
<th>Novice Visual</th>
<th>Expert Full</th>
<th>Expert Visual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of minutes spent per topic (fixed by study)</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Text queries issued per topic</td>
<td>9.04</td>
<td>14.33</td>
<td>4.33</td>
<td>5.21</td>
</tr>
<tr>
<td>Word count per text query</td>
<td>1.51</td>
<td>1.55</td>
<td>1.54</td>
<td>1.30</td>
</tr>
<tr>
<td>Number of video story segments returned by each text query</td>
<td>105.29</td>
<td>15.65</td>
<td>79.40</td>
<td>20.14</td>
</tr>
<tr>
<td>Image queries per topic</td>
<td>1.23</td>
<td>1.54</td>
<td>1.13</td>
<td>6.29</td>
</tr>
<tr>
<td>Precomputed feature sets (e.g., “roads”) browsed per topic</td>
<td>0.13</td>
<td>0.21</td>
<td>0.83</td>
<td>1.92</td>
</tr>
</tbody>
</table>
Breakdown, Origins of Submitted Shots

- **Expert Full**:  
  - Text query
  - Image query
  - Best-of set

- **Novice Full**:  
  - Text query
  - Image query

- **Expert Visual-Only**:  
  - Text query
  - Image query

- **Novice Visual-Only**:  
  - Text query
  - Image query
Breakdown, Origins of Correct Answer Shots

Expert Full

Novice Full

Expert Visual-Only

Novice Visual-Only

- text query
- image query
- best-of set
Manual and Automatic Search

• Use text retrieval to find the candidate shots
• Re-rank the candidate shots by linearly combining scores from multimodal features
  • Image similarity (color, edge, texture)
  • Semantic detectors (anchor, commercial, weather, sports...)
  • Face detection / recognition
• Re-ranking weights trained by logistic regression
  • Query-Specific-Weight
    - Trained on development set (truth collected within 15 min)
    - Training on pseudo-relevance feedback
  • Query-Type-Weight
    - 5 Q-Types: Person, Specific Object, General Object, Sports, Other
    - Trained using sample queries for each type
Text Only vs. Text & Multimodal Features

- Multimodal features are slightly helpful with weights trained by pseudo-relevance feedback.
- Weights trained on development set degrade the performance.
Development Set vs. Testing Set

- “Train-on-Testing” >> “Text only” > “Train-on-Development”
  - Multimodal features are helpful if the weights are well trained
  - Multimodal features with poorly trained weights hurt
  - Difference of data distribution b/w development and testing data

Carnegie Mellon
• Anchor is the most useful non-textual feature
• Face detection and recognition are slightly helpful
• Overall, image examples are not useful
Contributions of Non-Textual Features (by Topic)

- Face recognition – overall helpful
  - ++ “Hussein”, +++ “Donaldson”
  - - “Clinton”, “Hyde”, “Netanyahu”
- Face detection (binary) – overall helpful
  - + “golfer”, “people moving stretcher”, “handheld weapon”
- Anchor – overall & consistently helpful
  - + all person queries
- HSV Color – slightly harmful
  - ++ “golfer”, + “hockey rink”, + “people with dogs”
  - -- “Bicycle”, “umbrella”, “tennis”, “Donaldson”
Conclusions

• The relative high information retrieval performances by both experts and novices are due to reliance on an intelligent user possessing excellent visual perception skills to compensate for comparatively low precision in automatically classifying the visual contents of video.

• Visual-only *interactive* systems better than full-featured manual or automatic systems.

• ASR and CC text enable better interactive, manual, and automatic retrieval.

• Anchor and face improve manual/automatic search over just text.

• Novices will need additional interface scaffolding and support to try interfaces beyond traditional text search.
TRECVID 2004 Concept Classification

- **Boat/ship**: video of at least one boat, canoe, kayak, or ship of any type.
- **Madeleine Albright**: video of Madeleine Albright
- **Bill Clinton**: video of Bill Clinton
- **Train**: video of one or more trains, or railroad cars which are part of a train
- **Beach**: video of a beach with the water and the shore visible
- **Basket scored**: video of a basketball passing down through the hoop and into the net to score a basket - as part of a game or not
- **Airplane takeoff**: video of an airplane taking off, moving away from the viewer
- **People walking/running**: video of more than one person walking or running
- **Physical violence**: video of violent interaction between people and/or objects
- **Road**: video of part of a road, any size, paved or not
TRECVID 2004 Concept Classification

- **Boat/ship**: video of at least one boat, canoe, kayak, or ship of any type.
- **Madeleine Albright**: video of Madeleine Albright
- **Bill Clinton**: video of Bill Clinton
- **Train**: video of one or more trains, or railroad cars which are part of a train
- **Beach**: video of a beach with the water and the shore visible
- **Basket scored**: video of a basketball passing down through the hoop and into the net to score a basket - as part of a game or not
- **Airplane takeoff**: video of an airplane taking off, moving away from the viewer
- **People walking/running**: video of more than one person walking or running
- **Physical violence**: video of violent interaction between people and/or objects
- **Road**: video of part of a road, any size, paved or not
CAUTION: Changing MAP with users/topic

It is likely that MAP for a group can be trivially improved by merely adding more users/topic with a simple selection strategy.
Thank You

Carnegie Mellon University
## TRECVID 2004 Search Topics

<table>
<thead>
<tr>
<th>Type</th>
<th>Generic</th>
<th>Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objects</td>
<td>Buildings with flood waters, hockey net, umbrellas, wheelchairs</td>
<td>U.S. Capitol dome</td>
</tr>
<tr>
<td>People</td>
<td>Street scenes, people walking dogs, people moving stretcher, people going up/down steps, protest/march with signs</td>
<td>Henry Hyde, Saddam Hussein, Boris Yeltsin, Sam Donaldson, Benjamin Netanyahu, Bill Clinton with flag</td>
</tr>
<tr>
<td>Events</td>
<td>Fingers striking keyboard, golfer making shot, handheld weapon firing, moving bicycles, tennis player hitting ball, horses in motion</td>
<td></td>
</tr>
<tr>
<td>Scenes</td>
<td>Buildings on fire</td>
<td></td>
</tr>
</tbody>
</table>
TRECVID 2004 Example Images for Topics
Evaluation - TRECVID Search Categories

**AUTOMATIC:**
- **TOPIC** → **QUERY** → **SYSTEM** → **RESULT**
  - System directly evaluates query
  - System takes query as input and produces results without further human intervention

**MANUAL:**
- **TOPIC** → **HUMAN** → **QUERY** → **SYSTEM** → **RESULT**
  - Human formulates query based on topic and query interface, not on knowledge of collection or search results
  - System takes query as input and produces results without further human intervention

**INTERACTIVE:**
- **TOPIC** → **HUMAN** → **QUERY** → **SYSTEM** → **RESULT**
  - Human (re)formulates query based on topic, query, and/or results
  - System takes query as input and produces result without further human intervention on this invocation

---

**Evaluation - TRECVID Search Categories**

**Automatic:**
- System directly evaluates query
- System takes query as input and produces results without further human intervention

**Manual:**
- Human formulates query based on topic and query interface, not on knowledge of collection or search results
- System takes query as input and produces results without further human intervention

**Interactive:**
- Human (re)formulates query based on topic, query, and/or results
- System takes query as input and produces result without further human intervention on this invocation