Planning for TRECVID 2008

Discussion items from the planning session
at the TRECVID 2007 workshop

Data for 2008

❑ **More Sound and Vision**
  - ❑ ~100 hours for test
  - ❑ 2007 development + test for use as 2008 development
  - ❑ ASR (UTwente?)
  - ❑ MT (Queen Mary U. London?)

❑ **More BBC dramatic rushes**
  - ❑ ~50 videos from (~150 videos, ~60 hours)
  - ❑ Same BBC series

❑ **UK Airport surveillance**
  - ❑ 5 cameras x 20 hours

System tasks

- ❑ High-level feature extraction
- ❑ Search
- ❑ BBC rushes summarization
- ❑ Event detection in surveillance

High-level feature extraction

- ❑ Task
  - ❑ Choose features to evaluate that are more appropriate to S&V data?
- ❑ Training conditions
  - ❑ Retain A, B, C, a, b, c?
  - ❑ require a/b/c?
  - ❑ TV2007 data as development data for 2008
- ❑ Required runs
  - ❑ Retain Text-only?
  - ❑ Retain Visual-only?
- ❑ Evaluation
  - ❑ Normalize contribution of each feature's IAP to mean IAP so outliers have less effect? Use "median average precision"?
  - ❑ Report per feature random baseline?
  - ❑ Report per feature type or frequency class?
- ❑ Community annotation of additional S&V data
  - ❑ More detail in the feature "definitions"?
  - ❑ Using active learning?
  - ❑ Start with most frequent features in small sample?
  - ❑ Annotate region info? … only if appropriate for needed features
  - ❑ Two-month effort to document "dream system" based on lessons learned?
  - ❑ Some groups did detailed annotation for 2007 and will share? Tsinghua, UCSB, …?

Search

- ❑ Task
  - ❑ Collaborative allowed?
  - ❑ Special emphasis? (e.g. on events)
  - ❑ Reduce max search time to less than 15 mins – set limit based on feedback about how MAP increases with elapsed search time (data from LIG, others?)
  - ❑ Double topics for automatic systems by reusing 2007 topics + 2008? (Move to inferred AP for search)
- ❑ Training conditions
  - ❑ Retain A, B, C, a, b, c?
  - ❑ Require a/b/c ?
- ❑ Required runs
  - ❑ Retain text-only?
  - ❑ Retain visual-only?
- ❑ Evaluation
  - ❑ Community contribution to topics?
  - ❑ Community contribution to assessments?
  - ❑ Report MAP normalized relative to best performance for each topic?

Rushes summarization

- ❑ Task
    - ❑ Reduce upper limit on duration to 2% of original?
    - ❑ Allow two runs per group?
    - ❑ Require one run of fixed duration (2%)?
    - ❑ Require one run of duration <= 2%?
- ❑ Evaluation
    - ❑ Better way of assessing redundancy?
    - ❑ Better way of assessing ease of use?
        - ❑ More detailed questions:
            - ❑ "System has removed all junk" (strongly agree ..strongly disagree)
            - ❑ "Summary is 'smooth'"?
    - ❑ Provide more detailed ground truth?
        - ❑ Community adds detail after the evaluation?
    - ❑ Have systems provide start/stop times for each segment to be included in their summary … to be used for? Automatic eval? Automatic synthesis of simple summaries?
    - ❑ Distribute all the rest of the BBC rushes even if not all used in the evaluation… check with other Rushes projects

Event detection in surveillance discussion

- Task
  - What are the "natural" events that will be represented in the data?
  - Systems may be required to emit a confidence score for each detected event instance (TBD)
  - Will a possibly interactive, retrospective search task be supported?
- What will the training conditions be?
- Evaluation
  - LDC leading the initial annotation effort
  - Is community annotation of surveillance data feasible?
  - Metrics to be further developed with input from participants
  - Collaboration with other evaluation efforts possible in future
  - Teaming arrangements encouraged
    - Experimentation with Open Science Grid for data sharing, compute cycles, etc. being considered

Final notes:

❑ Please turn in your completed workshop survey

❑ Please send a copy of your slides to George Awad (George.Awad@nist.gov)

❑ The workshop papers and slides (as .pdf) will be posted on the website within the next month

❑ Final versions of the notebook papers are due to George.Awad@nist.gov by 1. March 2008