Event Model for Auto Video Search

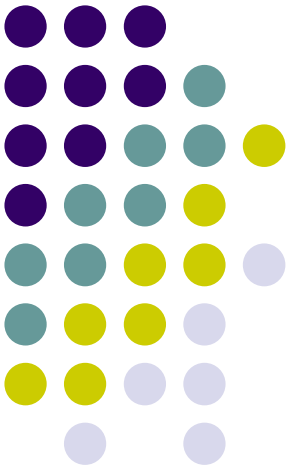
TRECVID 2005 Search by NUS PRIS

Tat-Seng Chua, Shi-Yong Neo, Hai-Kiat Goh, Ming Zhao, Yang Xiao & Gang Wang

(National University of Singapore)

Sheng Gao, Kai Chen, Qibin Sun & Qi Tian

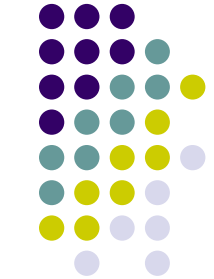
(Institute for Infocomm Research)



Emphasis of Last Year's System

- Query-dependent Model for retrieval
  - Uses query-class property to determine the parameters for fusion of various features
  - Effective in human, sports queries; not effective for more general queries as queries are heterogeneous
  - Provide a good basis for automatic fusion of various multimodal features (training using GMM)
- Use external resources for inducing query context
- Use of High level feature
  - Effectiveness of high level feature is limited as query requirements are generally different from high level features.

This Year's Emphasis-1



- Use of Event-based Entities for retrieval
  - makes use of the relevant external information collected from the web to generate domain knowledge in terms of timed-events
  - forms an important facet in retrieval and captures information that is not available in the text transcripts
  - *We recount earlier in previous talks by HLF teams that textual features plays a lesser role as they contains more error this year*

An Example from last year

- *Find shots that contain buildings covered in flood water*.

- *Disaster-type queries, event-oriented.*
- *Retreival can be done effectively if we know the flooding events, location, time, etc*
- *Such event information can be extracted online*

Examples from this year

- Find shots of Condoleeza Rice.
- Find shots of Iyad Allawi, the former prime minister of Iraq.









Examples from this year-cont'

- Multi-lingual news video corpus, non-English names like (*Mahmoud Abbas, Allawi Iyad*, etc) cannot be easily recognized or translated à high error rate
- Greatly affect the number of retrievable relevant shots especially when the person's name plays an important part
- With event information, we can make use of location and time to recover these missing shots à predict the presence of these people in the news stories.
- Locations are seldom misrecognized or wrongly translated even for spoken documents since they are not as vulnerable to errors as person's names.

This Year's emphasis-2

- Use of High Level features
  - Integrates the results from high level feature extraction task to support general queries

Car

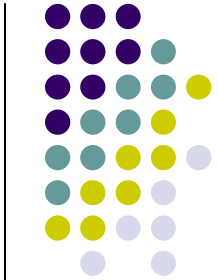
Explosion

Map

Using results from high level feature extraction task

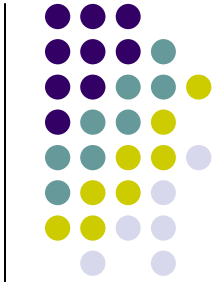
- Combining results from 21 participating groups using a rank-based fusion technique.
- 10 high level features available
  - Sports, car, explosion, maps, etc
- Extremely useful for answering general queries this year
- Useful for queries like: “Find shots of road with one or more cars”, “Find shots of tall building”, sports related queries

Main Presentation



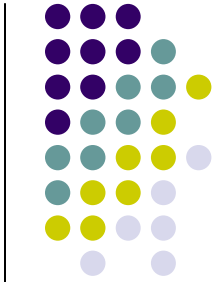
- Content Preprocessing
- Retrieval
- Result Analysis
- Conclusions

Content Preprocessing-1



- Automatic Speech Recognition and Machine Translated Text
  - Focus only on the English Text (Microsoft Beta) & Machine Translated English Text (Given by TRECVID)
    - Query in English
    - Our retrieval system is using only English lexical resources
  - Use phrase as base unit for analysis and retrieval
- Video OCR
  - By CMU
- Annotated High Level Features from High-level Feature Extraction Task
  - Next Slide

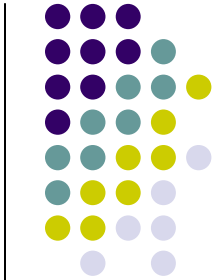
Content Preprocessing-2



- Annotated High Level Features from High-level Feature Extraction Task
  - 2 methods are used for combining various rank-lists:
    - Rank-based method (which is used in our Submitted runs):
      - Counting occurrences of a particular shot which is being ranked in the top 2000 shots by every group

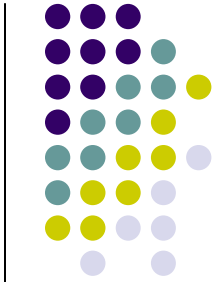
        minimum (Count($Shot_A$)> 6)
      - Score($Shot_A$) is given by averaging 4 of the most highly ranked positions (bias against shots which appears frequently but ranked lower)
      - MAP achievable 0.38 (slightly above best systems)
    - Rank-Boosting
      - Fuse the ranklist according to performance of various system, but only can be done when the performance is known or training data is available.
      - MAP achievable 0.44

Content Preprocessing-3



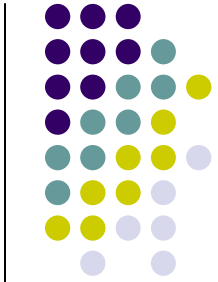
- Face Detection and Recognition
  - Based on Haar-like features
  - Recognition based on predefined set of 15 most commonly appearing Names (in ASR), (coincide with 3 human queries)
  - Face recognition on 2DHMM
- Audio Genre
  - cheering, explosion, silence, music, female speech, male speech, and noise.
- Shot Genre
  - sports, finance, weather, commercial, studio-anchor-person, general-face and general-non-face.
- Story Boundary
  - Donated results from IBM & Columbia U.

Content Preprocessing-4



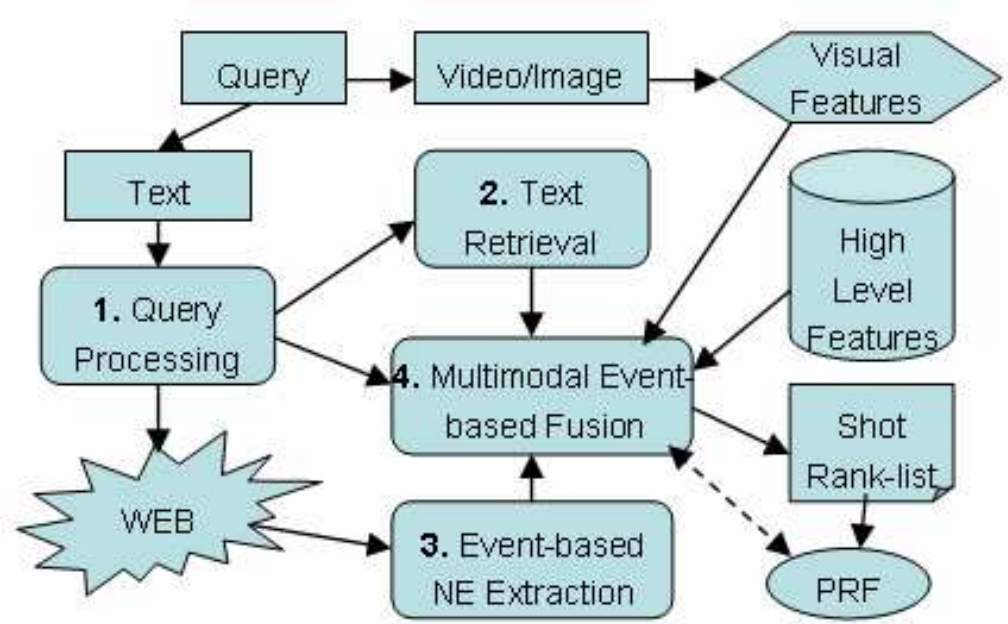
- Locality-temporal Information from News Video Stories
  - Mainly: Location, time, people
  - Based on stories boundaries (provided by IBM, Columbia U)
  - Person involved : Person's name who are mentioned within story ASR, MT.
  - Location of story
    - Iraq, Baghdad à choose Baghdad (more specific)
    - Normally mentioned right at the beginning of story
  - Time:
    - Video date, -1 day or -2 days
    - Cue terms à happened yesterday, this morning

Content Preprocessing-5



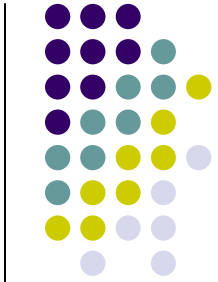
- Locality-temporal Information from News Video Stories
  - Story boundaries (hard to detect)
    - Good accuracies by IBM, Columbia U à around 75%
  - Location-type NEs and Time-type NEs
    - Tagging accuracy known to be over 90%
      - mildly affected by recognition and translation errors
    - Assigning story location or occurrence to news video story is found to be 82% based on part of training set.
  - Minimizing noise à discarding non useful segments (i.e. commercial news, led-in) segments longer than 200 seconds or less than 12 seconds

Retrieval



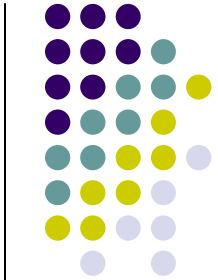
- 4 main stages:
  - query processing
  - text retrieval
  - event-based NE extraction from relevant online news articles
  - multimodal event-based fusion.

Retrieval- 2



- ***Query Processing***
  - Extracting keywords
  - Inducing query-class, {Person, Sports, Finance, Weather, Politics, Disaster and General}
  - Inducing explicit constraints.
  - Performing query expansion on parallel text corpus (based on high mutual information with the original query terms)
- ***ASR Retrieval***
  - ASR retrieval à vector-space model based on tf.idf score + %overlap with expanded words. More details found in our previous work **(**Chua et al, 2004**)**.

Retrieval -3



- ***Event-based NE Extraction from External News Sources***
  - Using the text query to retrieve related news articles (news corpus extracted online last year)
  - Performing morphological analysis on the related articles and then passed to the NE extractor module to obtain various NE types such as: Person Name, Location and Time.
  - Therefore, each news articles is been represented as a set of NEs denote by E'. While P' is set of NEs extracted from ASR/MT
  - Make use of a simple assumption to relate E' to P' by using NEs.

$$\mathrm{Re}l(P', E') = \sum_{i=Loc,\,Person,\,Time} \beta_i * Y(P'_i \cap E'_i)$$

  - where β_i is the weight given for different NE type, *Y* is the output number of intersections. Similarly, we can obtain the probability of *NE' or Event'* (given by query) relevance to a news video story in terms of location-time relation.

$$P(NE', Event' \mid P', E') = \alpha_m \,\mathrm{Re}\, l(P', E')$$

  - where α_m are weights given to different query types.

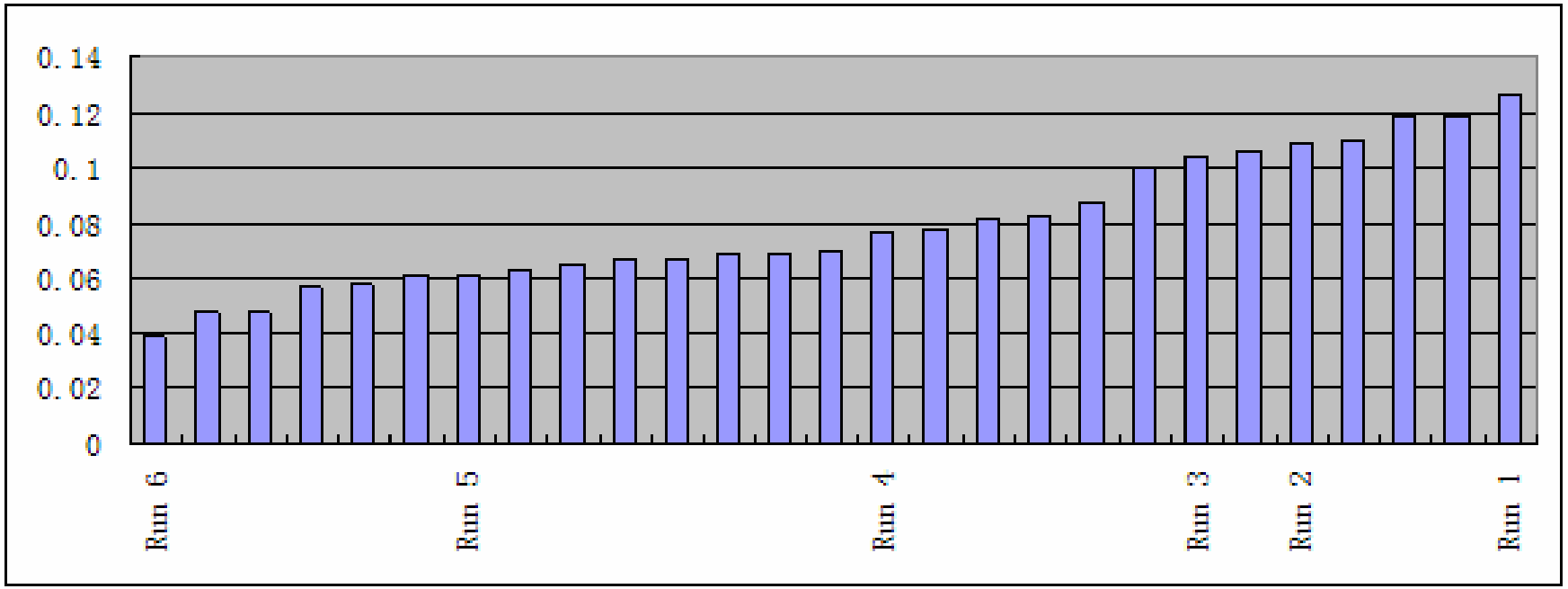
Retrieval -4

- ***Multimodal Fusion***
  - Different queries may have very different characteristics and hence require very different feature combinations
  - Uses a combination of heuristic weights, and the visual information obtained from the sample shots given to form an initial set of fusion parameters for the queries.
  - Subsequently perform a round of pseudo relevant feedback (PRF). This is done by using the top 20 return shots from each query.

Result Analysis

- We submitted a total of 6 runs
  - **Run 5.** (The required text-only run). The number of keywords in this case is restricted to 4. Using these keywords, we perform a basic retrieval on the ASR and MT using standard tf-idf function to obtain a ranked list of “phrases”.
  - **Run 4.** (Including other text). Run 4 is also a text-only run. The difference between Run 4 and Run 5 is the use of additional expanded words and context.
  - **Run 3.** (Run 4 with high level features). The weights of the shots are boosted in the following manners: (a) if the shot contains the high level feature that is found in the text query; and (b) if the shot contains the high level feature that is found in the given 6 sample videos.
  - **Run 2.** (Multimodal Run with Pseudo Relevance Feedback (PRF)). This run makes use of the various multimodal features extracted from the video to re-rank shots obtained in Run 3. Using the query-class information derive from the text query, weights are assigned to various multimodal features, similar to previous work in (Chua et al, 2004). (Type B)
  - **Run 1.** (Multimodal Event-based Run with PRF). This run makes use of all multimodal features in Run 2 as well as the fusion with an additional event entity feature (Neo et al, 2006). (Type B)
  - **Run 6.** (Visual only). This run uses only visual features. The purpose of this run is to test the underlying retrieval result if all textual features are discarded.

Result Analysis -2



Run 5. (The required text-only run).
Run 4. (Including other text).
Run 3. (Run 4 with high level features).
Run 2. (Multimodal Run with Pseudo Relevance Feedback (PRF)).
Run 1. (Multimodal Event-based Run with PRF).
Run 6. (Visual only).

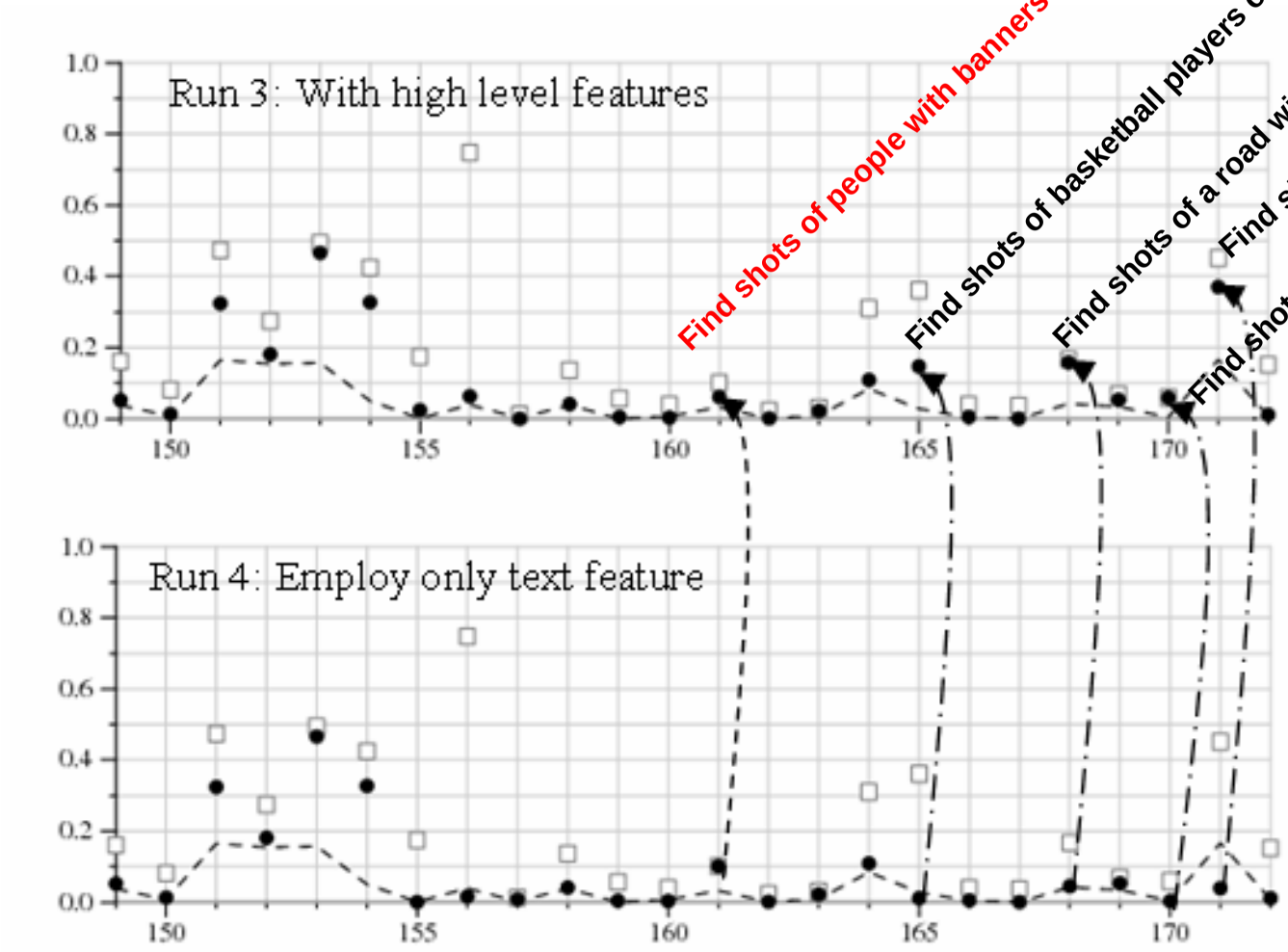
Result Analysis -3

30% improvement

	Run1	Run2	Run3	Run4	Run5	Run6
MAP	0.126	0.109	0.104	0.076	0.061	0.039

Run 5. (The required text-only run).
Run 4. (Including other text).
Run 3. (Run 4 with high level features).
Run 2. (Multimodal Run with Pseudo Relevance Feedback (PRF)).
Run 1. (Multimodal Event-based Run with PRF).
Run 6. (Visual only).

Result Analysis -4



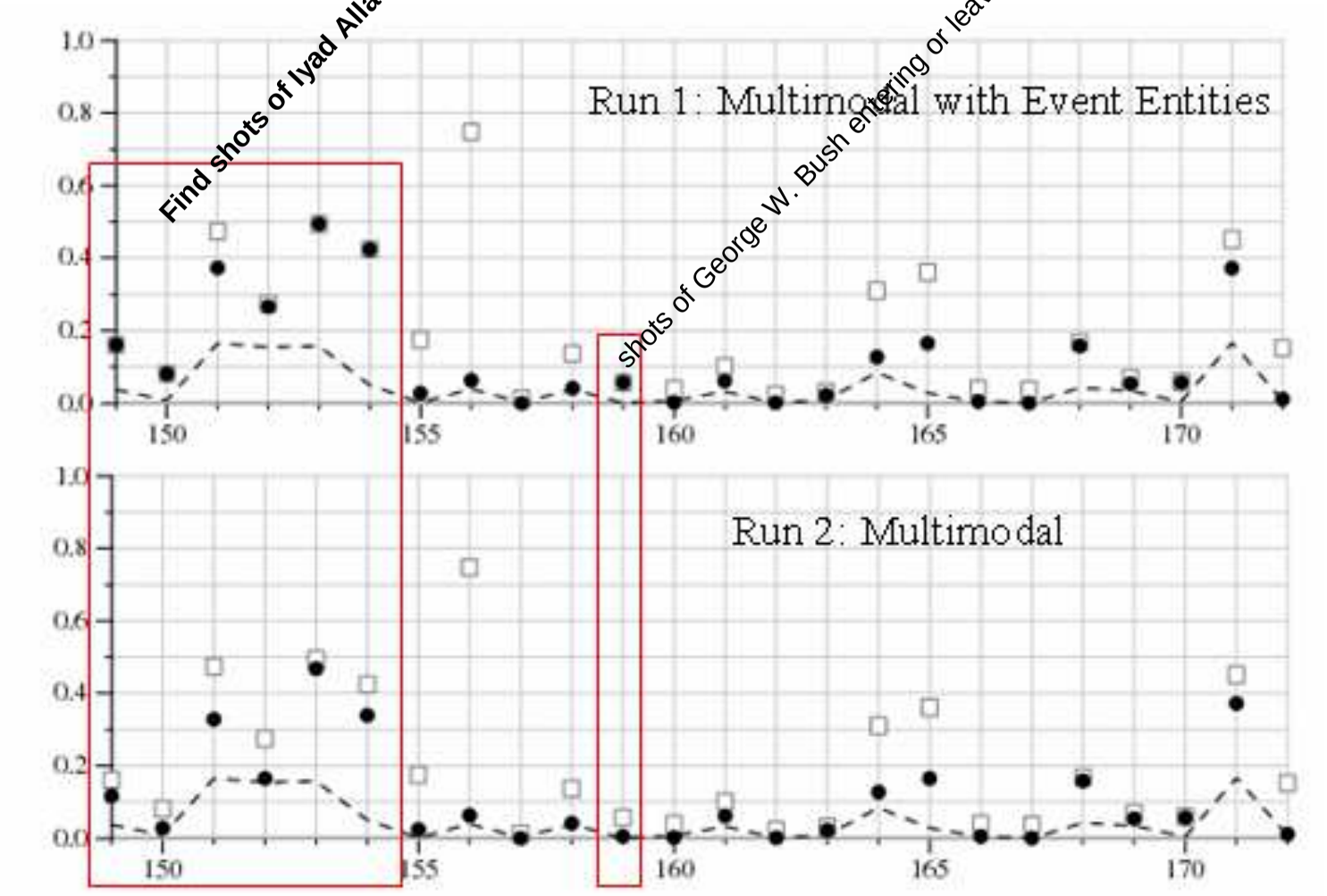
Result Analysis -3

15% improvement

	Run1	Run2	Run3	Run4	Run5	Run6
MAP	0.126	0.109	0.104	0.076	0.061	0.039

Run 5. (The required text-only run).
Run 4. (Including other text).
Run 3. (Run 4 with high level features).
Run 2. (Multimodal Run with Pseudo Relevance Feedback (PRF)).
Run 1. (Multimodal Event-based Run with PRF).
Run 6. (Visual only).

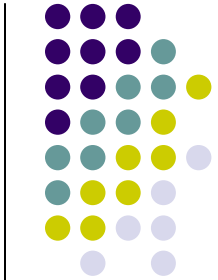
Result Analysis -5



Main Observations and conclusions

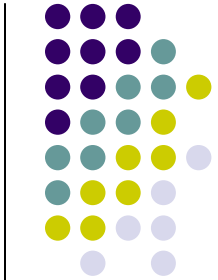
- Structural use of external resource (like event entities from news articles) is useful in supporting retrieval especially for NE queries and event queries
- High level features are useful for general queries this year
- Use of query-dependent retrieval as a basis for initial fusion is effective

Future works



- Improvement to matching between news video stories and external news articles
  - Currently using only Name Entities
  - Can include other type of pre-defined structures in news (i.e. sports, weather, etc…)
  - Or undefined structure by clustering various events?
- Providing better query-class by clustering and finding query-classes automatically
  - As seen in MM’05 (done by Columbia University)

End of Presentation



- Special thanks to all groups which have contributed valuable donated features…
- Questions are welcome.