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

1. Query-dependent Model for retrieval
   1. Uses query-class property to determine the parameters for fusion of various features
   1. Effective in human, sports queries; not effective for more general queries as queries are heterogeneous
   1. Provide a good basis for automatic fusion of various multimodal features (training using GMM)

1. Use external resources for inducing query context

1. Use of High level feature
   1. Effectiveness of high level feature is limited as query requirements are generally different from high level features.
This Year’s Emphasis-1

1. Use of Event-based Entities for retrieval
   1. makes use of the relevant external information collected from the web to generate domain knowledge in terms of timed-events
   1. forms an important facet in retrieval and captures information that is not available in the text transcripts
   1. *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

1. Find shots that contain buildings covered in flood water.

1. Disaster-type queries, event-oriented.

1. Retrieval can be done effectively if we know the flooding events, location, time, etc.

1. Such event information can be extracted online
Examples from this year

1. Find shots of Condoleezza Rice.
2. Find shots of Iyad Allawi, the former prime minister of Iraq.
Multi-lingual news video corpus, non-English names like *(Mahmoud Abbas, Allawi Iyad, etc)* cannot be easily recognized or translated → 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 → 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

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

1. Combining results from 21 participating groups using a rank-based fusion technique.
2. 10 high level features available
   1. Sports, car, explosion, maps, etc
3. Extremely useful for answering general queries this year
4. Useful for queries like: “Find shots of road with one or more cars”, “Find shots of tall building”, sports related queries
Main Presentation

1. Content Preprocessing
2. Retrieval
3. Result Analysis
4. Conclusions
Content Preprocessing-1

1. Automatic Speech Recognition and Machine Translated Text
   1. Focus only on the English Text (Microsoft Beta) & Machine Translated English Text (Given by TRECVID)
      1. Query in English
      1. Our retrieval system is using only English lexical resources
   1. Use phrase as base unit for analysis and retrieval

1. Video OCR
   1. By CMU

1. Annotated High Level Features from High-level Feature Extraction Task
   1. Next Slide
Content Preprocessing-2

1. Annotated High Level Features from High-level Feature Extraction Task
   2. 2 methods are used for combining various rank-lists:
      1. Rank-based method (which is used in our Submitted runs):
         $\checkmark$ Counting occurrences of a particular shot which is being ranked in the top 2000 shots by every group minimum (Count(Shot_A) > 6)
         $\checkmark$ Score(Shot_A) is given by averaging 4 of the most highly ranked positions (bias against shots which appears frequently but ranked lower)
         $\checkmark$ MAP achievable 0.38 (slightly above best systems)
      1. Rank-Boosting
         $\checkmark$ Fuse the ranklist according to performance of various system, but only can be done when the performance is known or training data is available.
         $\checkmark$ MAP achievable 0.44
Content Preprocessing-3

1. Face Detection and Recognition
   1. Based on Haar-like features
   2. Recognition based on predefined set of 15 most commonly appearing Names (in ASR), (coincide with 3 human queries)
   3. Face recognition on 2DHMM

1. Audio Genre
   1. cheering, explosion, silence, music, female speech, male speech, and noise.

1. Shot Genre
   1. sports, finance, weather, commercial, studio-anchor-person, general-face and general-non-face.

1. Story Boundary
   1. Donated results from IBM & Columbia U.
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 choose Baghdad (more specific)
- Normally mentioned right at the beginning of story

Time:

- Video date, -1 day or -2 days
- Cue terms happened yesterday, this morning
1 Locality-temporal Information from News Video Stories
   1 Story boundaries (hard to detect)
      1 Good accuracies by IBM, Columbia U around 75%
   1 Location-type NEs and Time-type NEs
      1 Tagging accuracy known to be over 90%
         $ mildly affected by recognition and translation errors
      1 Assigning story location or occurrence to news video story
         is found to be 82% based on part of training set.
   1 Minimizing noise discarding non useful segments (i.e.
      commercial news, led-in) segments longer than 200
      seconds or less than 12 seconds
Retrieval

1. 4 main stages:
   1. query processing
   2. text retrieval
   3. event-based NE extraction from relevant online news articles
   4. multimodal event-based fusion.
1 **Query Processing**

1. Extracting keywords
1. Inducing query-class, \{Person, Sports, Finance, Weather, Politics, Disaster and General\}
1. Inducing explicit constraints.
1. Performing query expansion on parallel text corpus (based on high mutual information with the original query terms)

1 **ASR Retrieval**

1. ASR retrieval vector-space model based on tf.idf score + %overlap with expanded words. More details found in our previous work (Chua et al, 2004).
**Event-based NE Extraction from External News Sources**

1. Using the text query to retrieve related news articles (news corpus extracted online last year)
2. 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.
3. 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
4. Make use of a simple assumption to relate $E'$ to $P'$ by using NEs.

\[
\text{Rel}(P', E') = \sum_{i=\text{Loc, Person, Time}} \beta_i * Y(P'_i \cap E'_i)
\]

where $\beta_i$ is the weight given for different NE type, $Y$ is the output number of intersections. Similarly, we can obtain the probability of \textit{NE’ or Event’} (given by query) relevance to a news video story in terms of location-time relation.

\[
P(\text{NE’}, \text{Event’} \mid P', E') = \alpha_m \text{Rel}(P', E')
\]

where $\alpha_m$ are weights given to different query types.
**Multimodal Fusion**

1. Different queries may have very different characteristics and hence require very different feature combinations.
2. 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.
3. Subsequently perform a round of pseudo relevant feedback (PRF). This is done by using the top 20 return shots from each query.
Result Analysis

1. We submitted a total of 6 runs
   1. **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”.
   1. **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.
   1. **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.
   1. **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)*
   1. **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)*
   1. **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

<table>
<thead>
<tr>
<th></th>
<th>Run1</th>
<th>Run2</th>
<th>Run3</th>
<th>Run4</th>
<th>Run5</th>
<th>Run6</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.126</td>
<td>0.109</td>
<td>0.104</td>
<td>0.076</td>
<td>0.061</td>
<td>0.039</td>
</tr>
</tbody>
</table>

- **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).

*30% improvement*
Result Analysis -4

Run 3: With high level features

Run 4: Employ only text feature

Find shots of people with banners or signs
Find shots of basketball players on the court
Find shots of a road with one or more cars
Find shots of a goal being made in a soccer match
Find shots of a tall building
**Result Analysis -3**

15% improvement

<table>
<thead>
<tr>
<th></th>
<th>Run1</th>
<th>Run2</th>
<th>Run3</th>
<th>Run4</th>
<th>Run5</th>
<th>Run6</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.126</td>
<td>0.109</td>
<td>0.104</td>
<td>0.076</td>
<td>0.061</td>
<td>0.039</td>
</tr>
</tbody>
</table>

- **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

Run 1: Multimodal with Event Entities

Run 2: Multimodal
Main Observations and conclusions

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

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

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