Combining Information Sources for Video Retrieval

Lowlands Team

Thijs Westerveld

Tzvetanka I. Ianeva

Liudmila Boldareva

Arjen P. de Vries

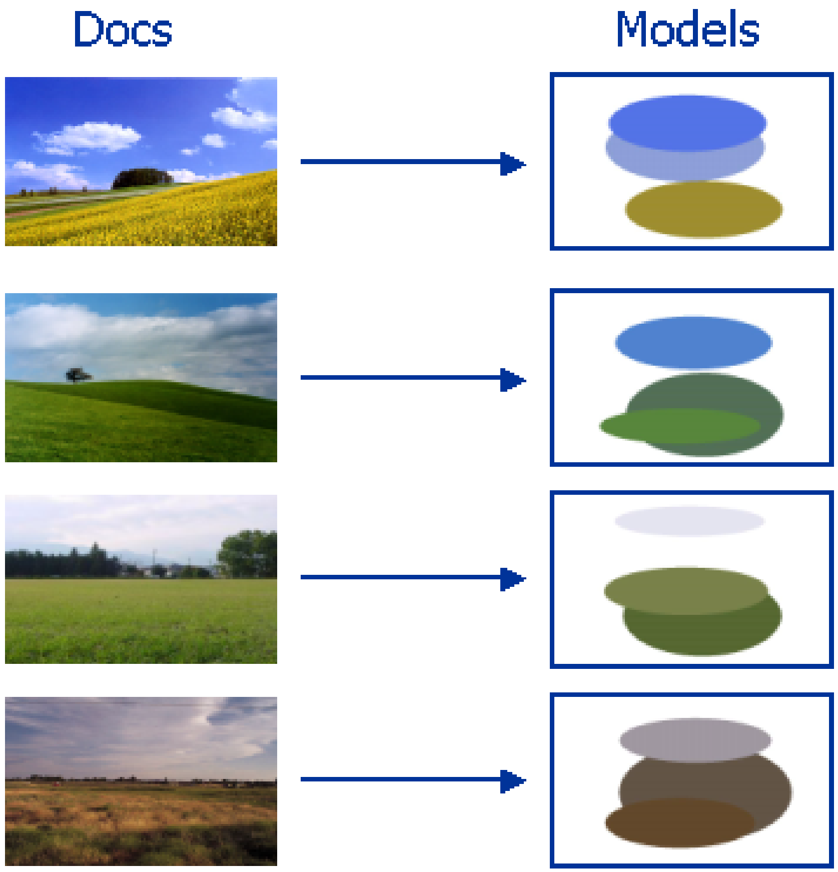
Djoerd Hiemstra

TRECVID workshop, 18 November 2003

Introduction

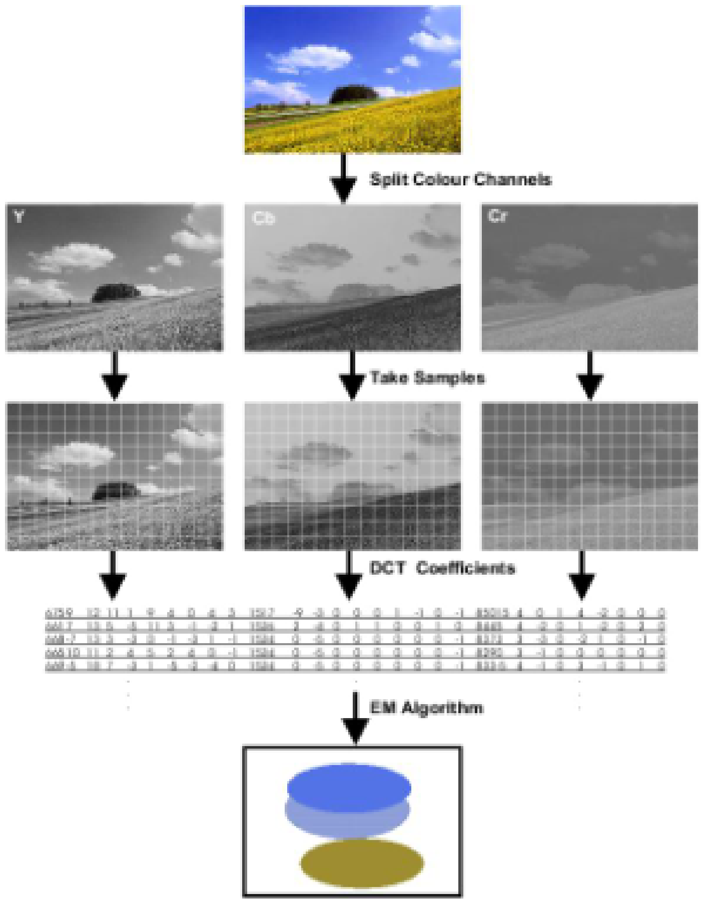
- Video Retrieval should take advantage of information from all available sources and modalities
  - ...but so far ASR best for almost any query
- LL11@TRECVID2003: Combining information sources
  - Different models/modalities
  - Multiple example images
  - Model similarity and human-judged similarity

Generative Probabilistic Model



- Indexing
  - Estimate Gaussian Mixture Models from images using EM
  - Based on feature vector with colour, texture and position information from pixel blocks
  - Fixed number of components

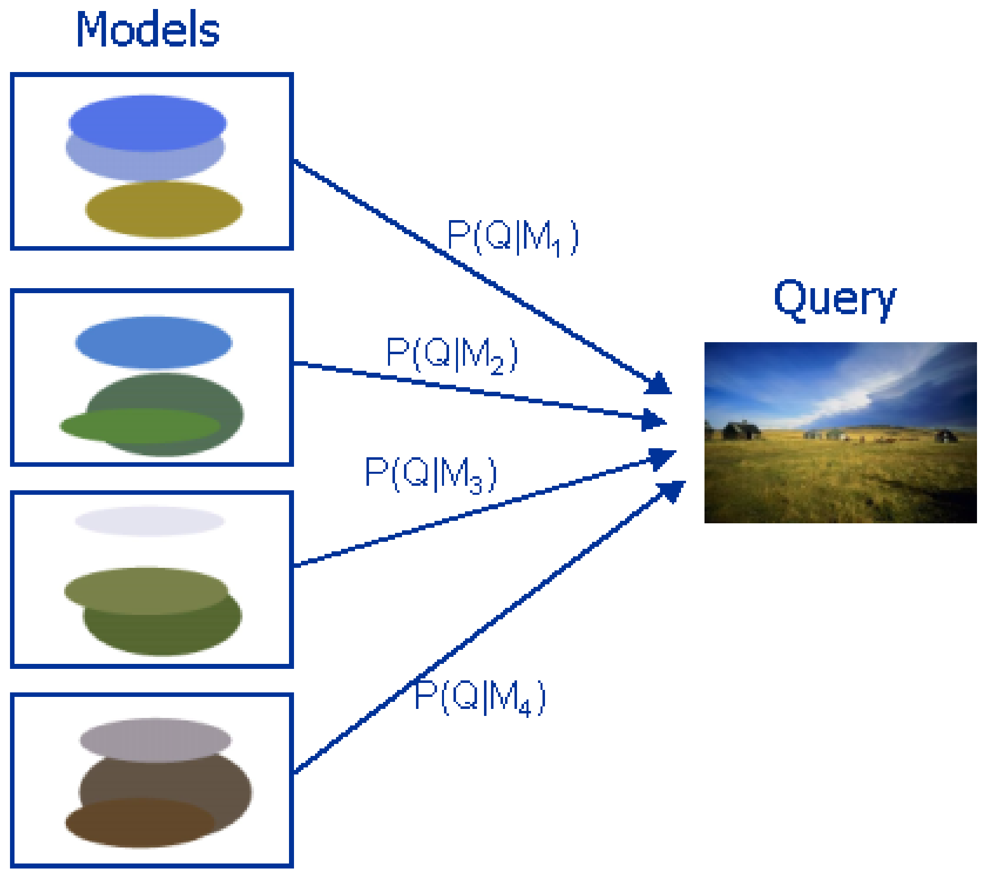
Generative Probabilistic Model



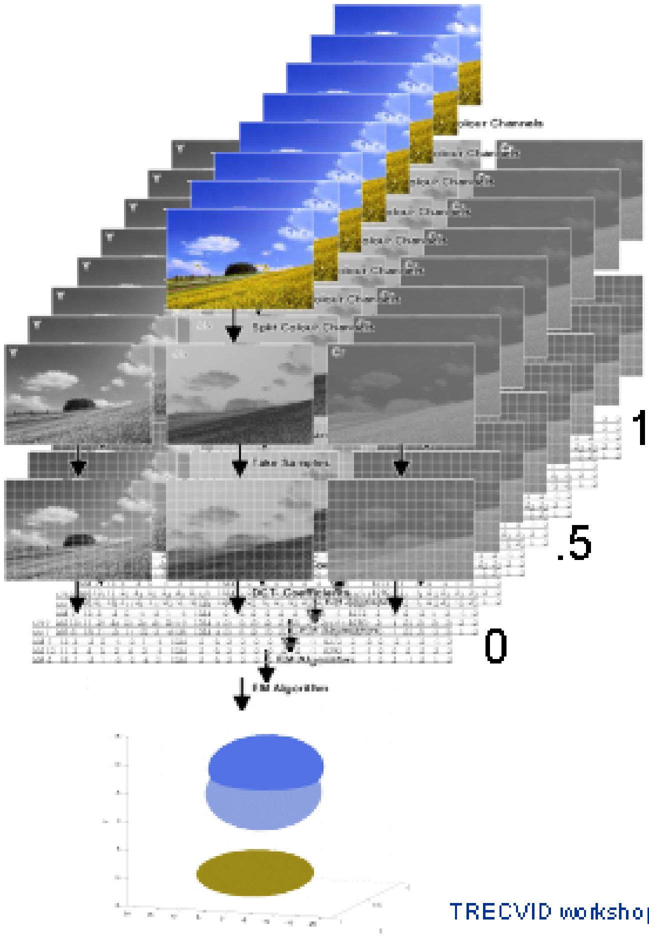
- Indexing
  - Estimate Gaussian Mixture Models from images using EM
  - Based on feature vector with colour, texture and position information from pixel blocks
  - Fixed number of components

Generative Probabilistic Model

- Retrieval
  – Calculate conditional probabilities of query samples given models in collection

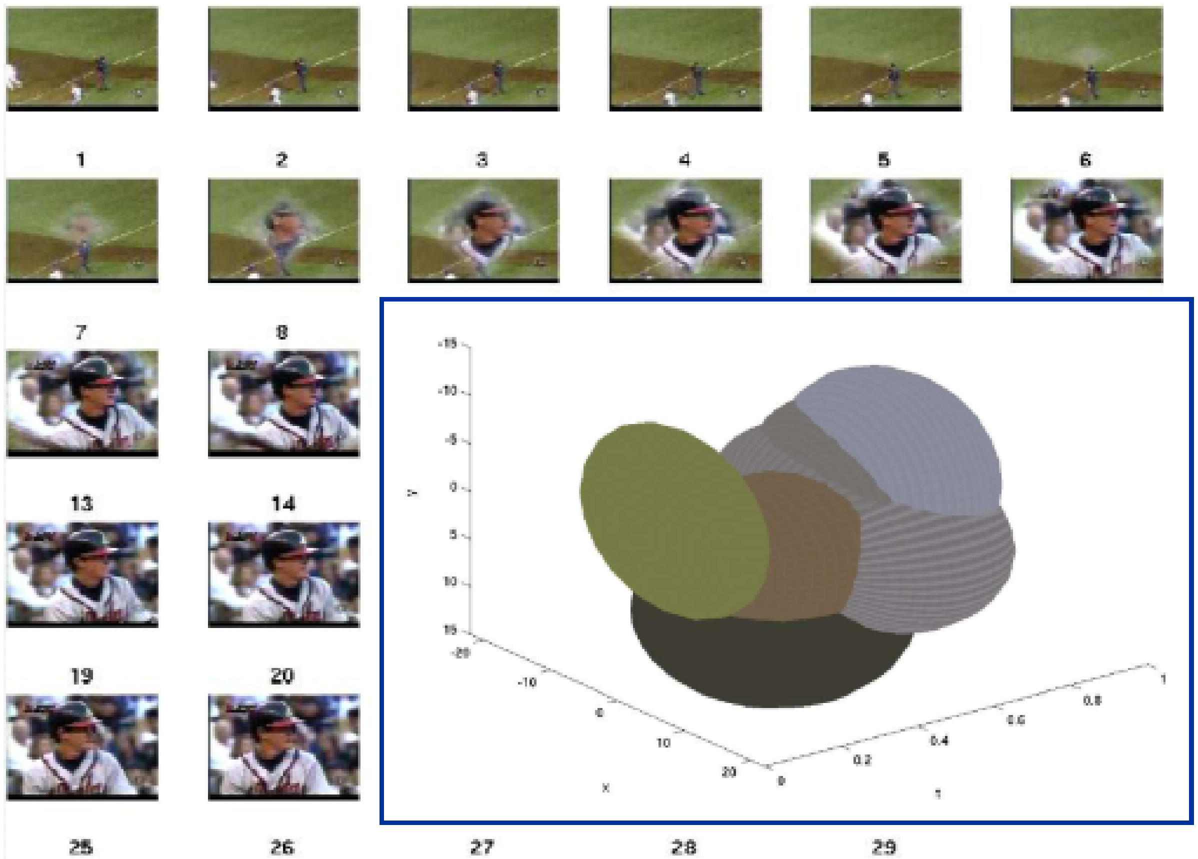


Dynamic Model



- GMM from multiple frames around KF
  - Feature vectors for each frame
  - Add time info
  - EM
- Dynamic model capture spatio-temporal information

Dynamic Model



Experimental Set-up

- Build models for each shot
  - Static, Dynamic, Language
- Build Queries from topics
  - Construct simple keyword text query
  - Select visual example
  - Rescale and compress example images to match video size and quality

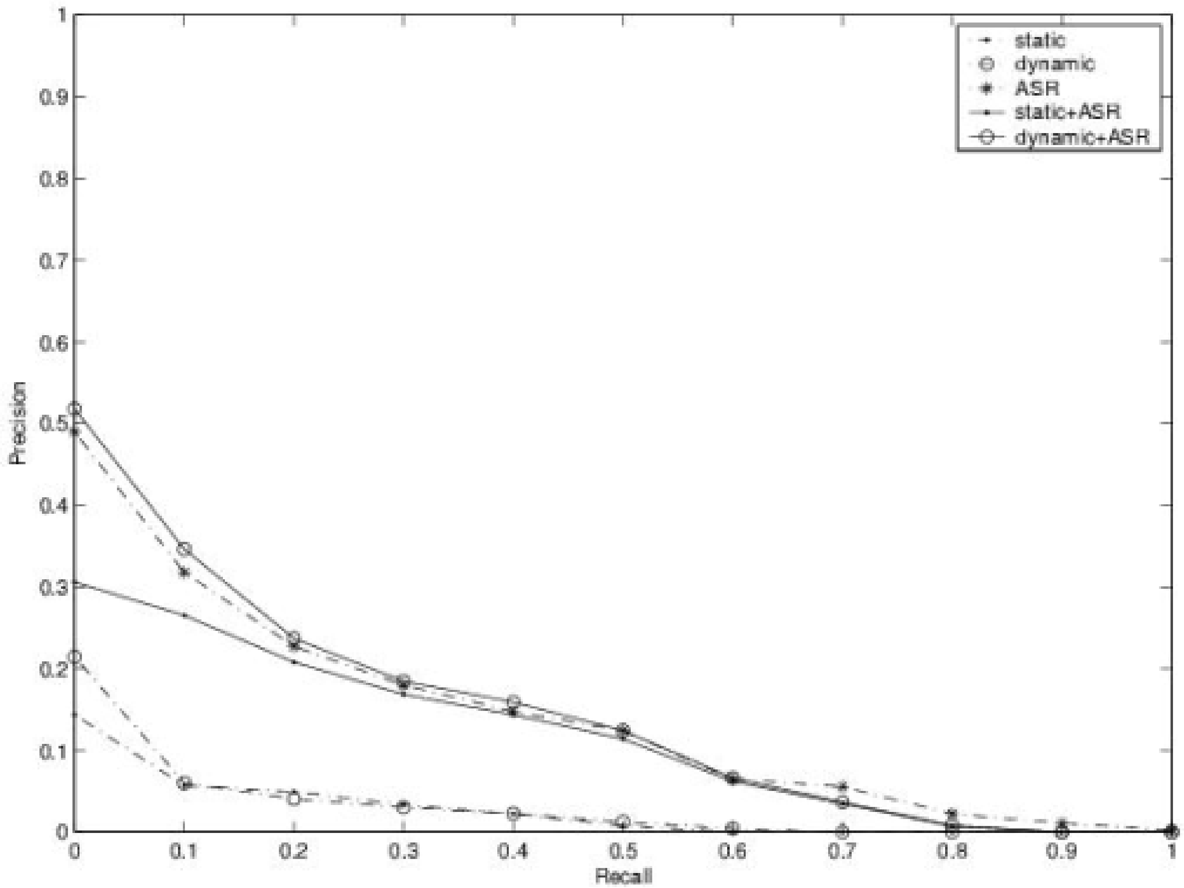
Combining Modalities

- Independence assumption textual/visual
  - $P(Q_t, Q_v|Shot) = P(Q_t|LM) * P(Q_v|GMM)$
- Strategy works well if both runs useful
  [CWI:TREC:2002]

- Dynamic run useful
- Static run not

Run	MAP
ASR only	.130
Static only	.022
Static+ASR	.105
Dynamic only	.022
Dynamic+ASR	.132

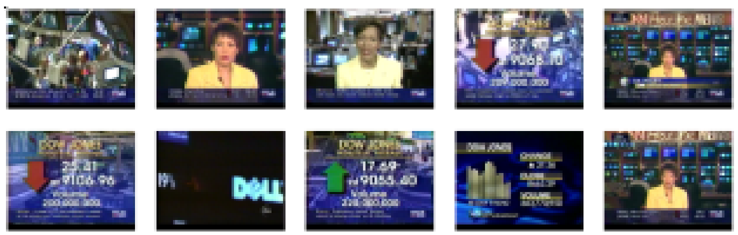
Combining Modalities



Combining Modalities

- Textual

"Dow Jones Industrial Average rise day points"



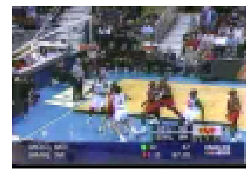
+

- Visual

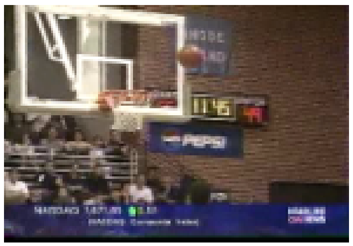
=

Merging Run Results

- Combining (conflicting) examples difficult
  [CWI:TREC:2002]
- Single example $\longrightarrow$ Miss relevant shots
- Round-Robin Merging
- MAP
  - All examples: .031
  - Combination .041
  - Single example: .022

		Combined
1	1	1
2	2	1
3	3	2
4	4	~~2~~
5	5	3
6	6	3
7	7	~~4~~
8	8	4
9	9	.
10	10	.

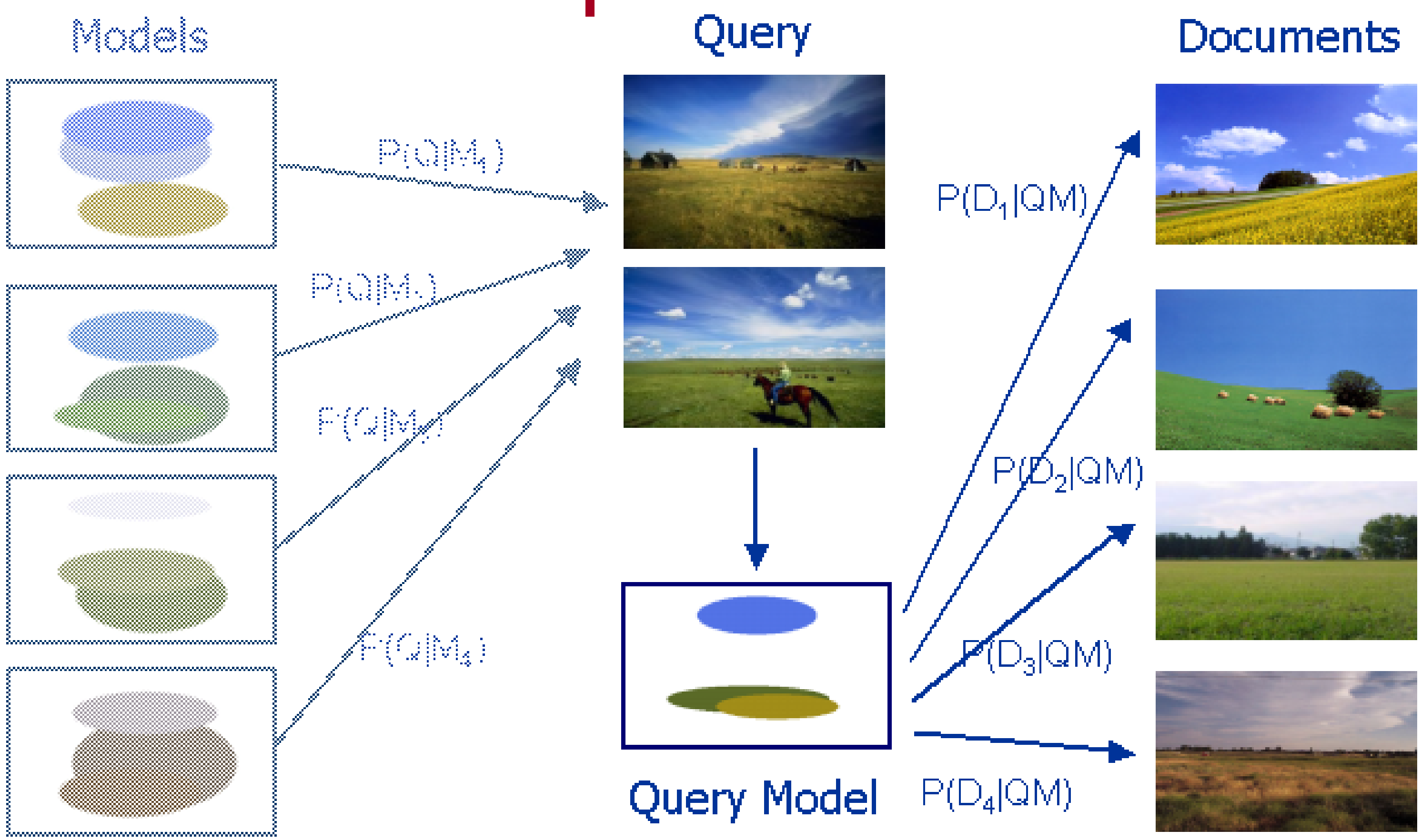
Merging Run Results



Top 30 Results

Top 30 Results

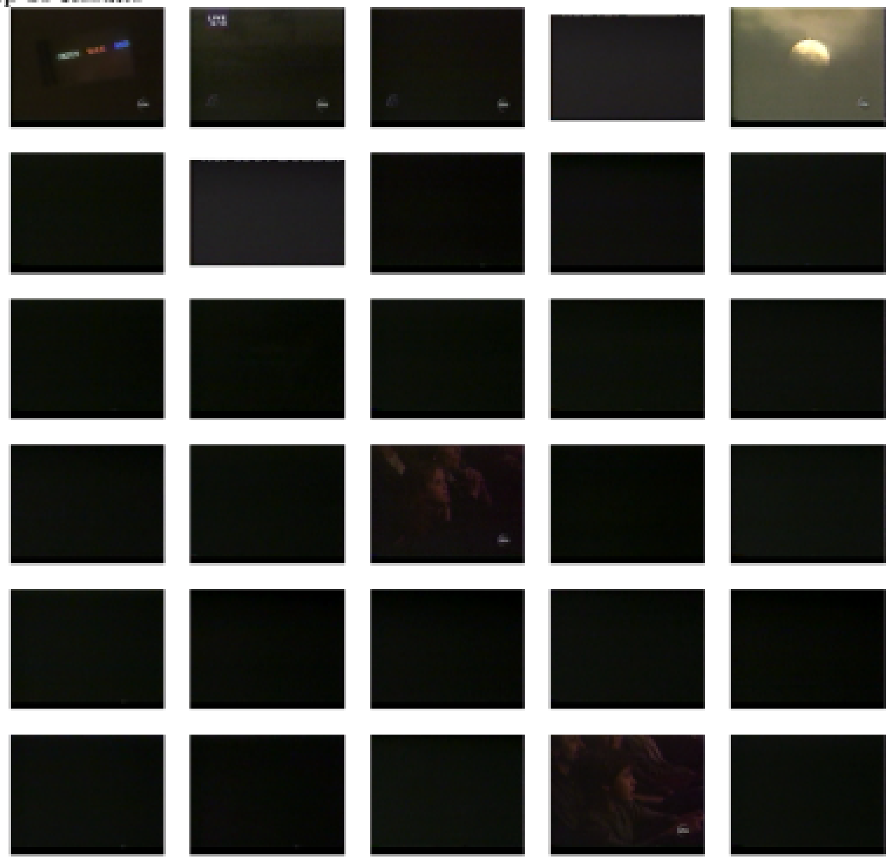
Topic Models



Topic Models

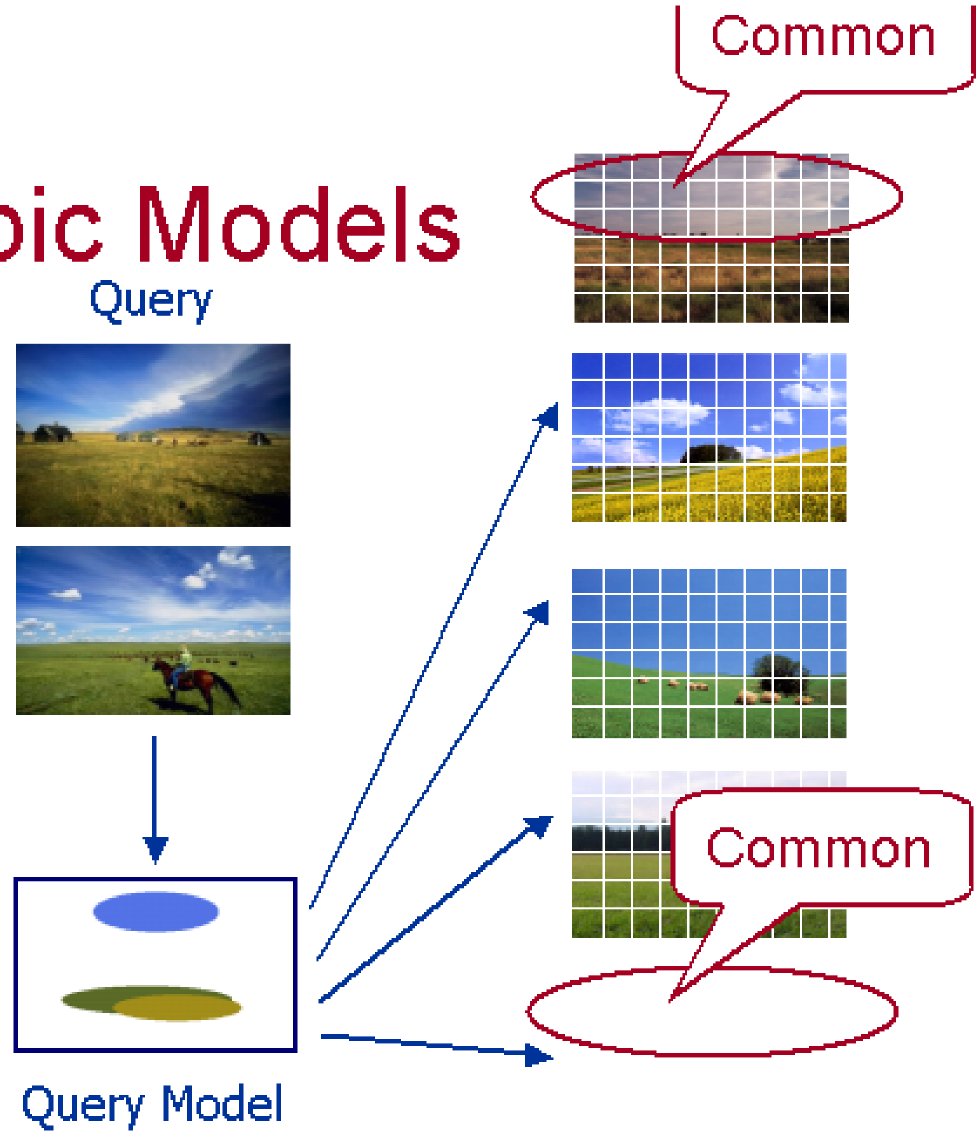
- Disappointing results
- Problems with common samples

Top 30 Results

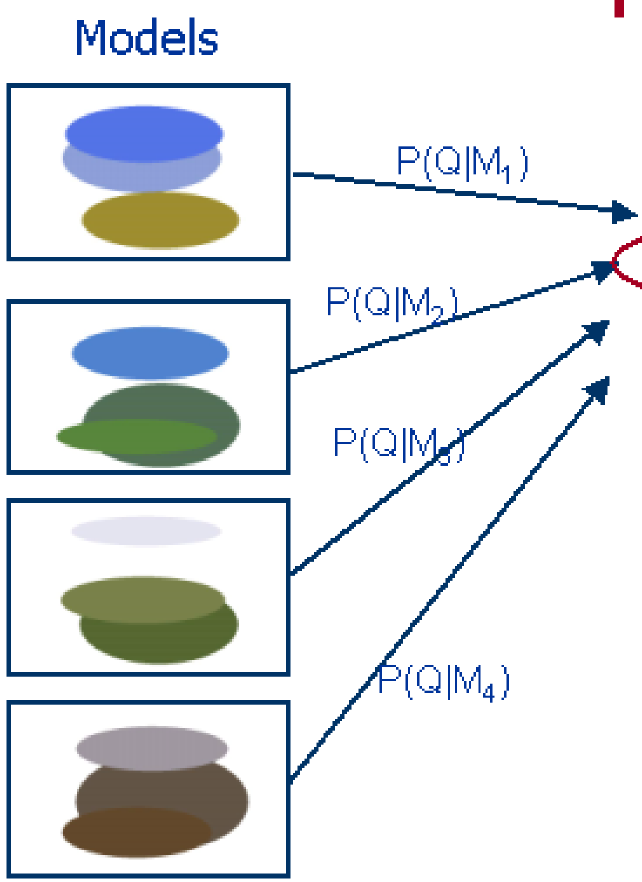


Topic Models

- Disappointing results
- Problems with common samples



Topic Models



- In original Retrieval setting common samples are no problem
- They get high scores from all document models
- → Does not influence ranking

Manual Conclusions

- Dynamic model captures temporal aspects
  - Also more training data, less dependent key-frame
- ASR+dynamic better than either alone
- Round-robin run combination useful
- Topic models have problems with common samples

Interactive Retrieval (data organisation)

- Pre-compute $\mathrm{P}(x|\mathrm{M})$ for all pairs
  - Use random sampling to build static GMM models to compute probabilities
- Only nearest neighbours are kept
- For the rest $\mathrm{P}(x|\mathrm{M}) = p^* = const$ is assumed
  - Reminder: $\mathrm{P}(x|\mathrm{M})$ is the probability that the query (x) was generated by the model M of the image that the user is interested in.

– "Trimming" allows fast interaction

– Reduces "noise" effect

Interactive Retrieval (algorithm)

- Begin ranking with text (topic + ASR)
- Positive feedback is added to the query ($\mathbf{x}$)
- Ranking scores come from probabilities and updated as:

$$P_{new}(M) = P_{old}(M) \cdot P(x_1 \ldots x_k | M) = P_{old}(M) \cdot \prod_k P(x_k | M)$$

(conditional independence assumption)

$x_1, \ldots x_k$ – positive feedback

- For next screen take the best ranked candidates according to P(M)

Interactive Experiments

- **3 + 1 Runs :**
  1. Random screen
  2. Text+ASR pre-formed sequence of screens

  ignored user input

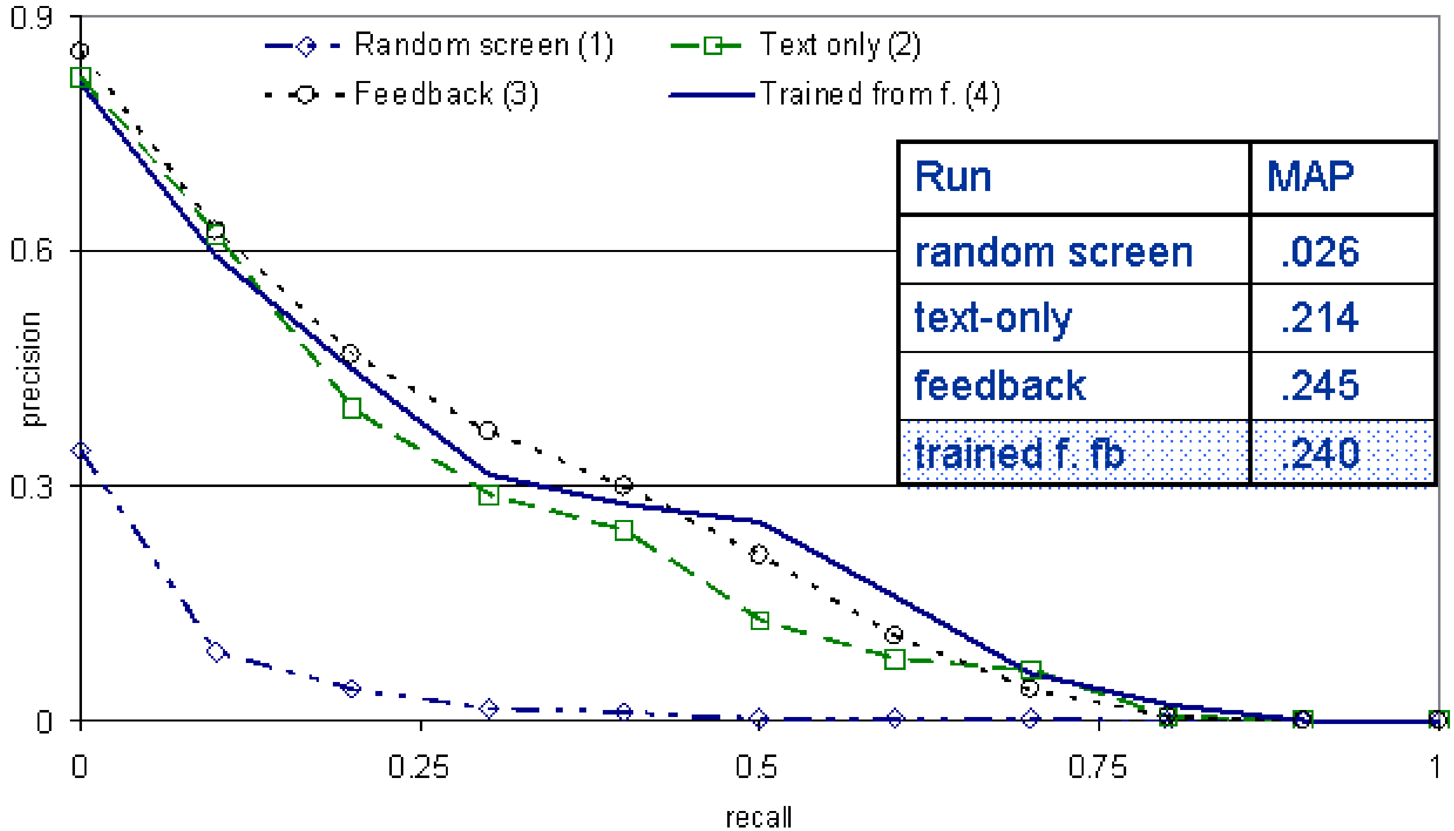
  3. Feedback updates screens
  4. idem, P(x|M) trained from (1)-(3)

  1 unofficial run

Browsing User Interface

- 4 x 3 frames
- Zoom possible
- Recent feedback shown
- Show / hide top-100 ranked
- Show / hide current topic

Interactive Experiments



Run	MAP
random screen	.026
text-only	.214
feedback	.245
trained f. fb	.240

In addition to MAP...

The following measures are meaningful:

– Amount of positive feedback given by the user.
  (if he/she is not stupid)
– Relevant shots produced by retrieval model
– Agreement between the user and NIST committee.
– How many displayed relevant shots the user missed.

Inside of iterations – feedback

Run	avg. % pos. judgments	found rel. vs. all submitted rel.(xx)
Random screen	0.41%	5.79% (380)
Text only	0.95%	37.83% (600)
Feedback	1.89%	49.18% (610)
Trained	4.35%	60.65% (681)

- More relevant frames on the screen, and earlier, in the systems with feedback
  - *User saw more "good" frames – user liked it!*
- The system gives good ranking even with small amount of feedback
  - *Text in prior ranking is important, and exploring Visual similarities helps!*

Inside of iterations – agreement

Run	agreement NIST & user	clicked vs.. displayed
Random screen	55.00%	68.75%
Text only	85.02%	48.09%
Feedback	78.74%	51.02%
Trained	64.03%	67.93%

- User often selects "best of worst" (see Random)
- Many missed relevant shots
  - Lost among other relevant shots (see Text only)
  - Key frames are not **KEY** frames

Inside of iterations – users

"Fast searching is better than slow"	"Annoying repeated images on screen"
"More relevant results on screen is encouraging"	Sometimes few or none good images: "I know there is another sphinx there"
Easy interface, image selection and zoom	"Some topic descriptions are vague"

- Users made many iterations (sometimes 150, average 65)
- Relevant shots popped up also at later time
- Fast interaction difficult with real videos… improve key frames!

Interactive Conclusions

- The role of user is important even with advanced techniques for similarity search
- Text gives a good start for interactive browsing
- Using visual features' nearest neighbours helps further
- Key frames are not enough: more sophisticated (re)presentation is needed
  - dynamic models / shot presentation