IBM Research TRECVID-2007 Video Retrieval System

Apostol (Paul) Natsev
IBM T. J. Watson Research Center
Hawthorne, NY 10532 USA

On Behalf Of:
Murray Campbell, Alexander Haubold, John R. Smith, Jelena Tesic, Lexing Xie, Rong Yan

Acknowledgement

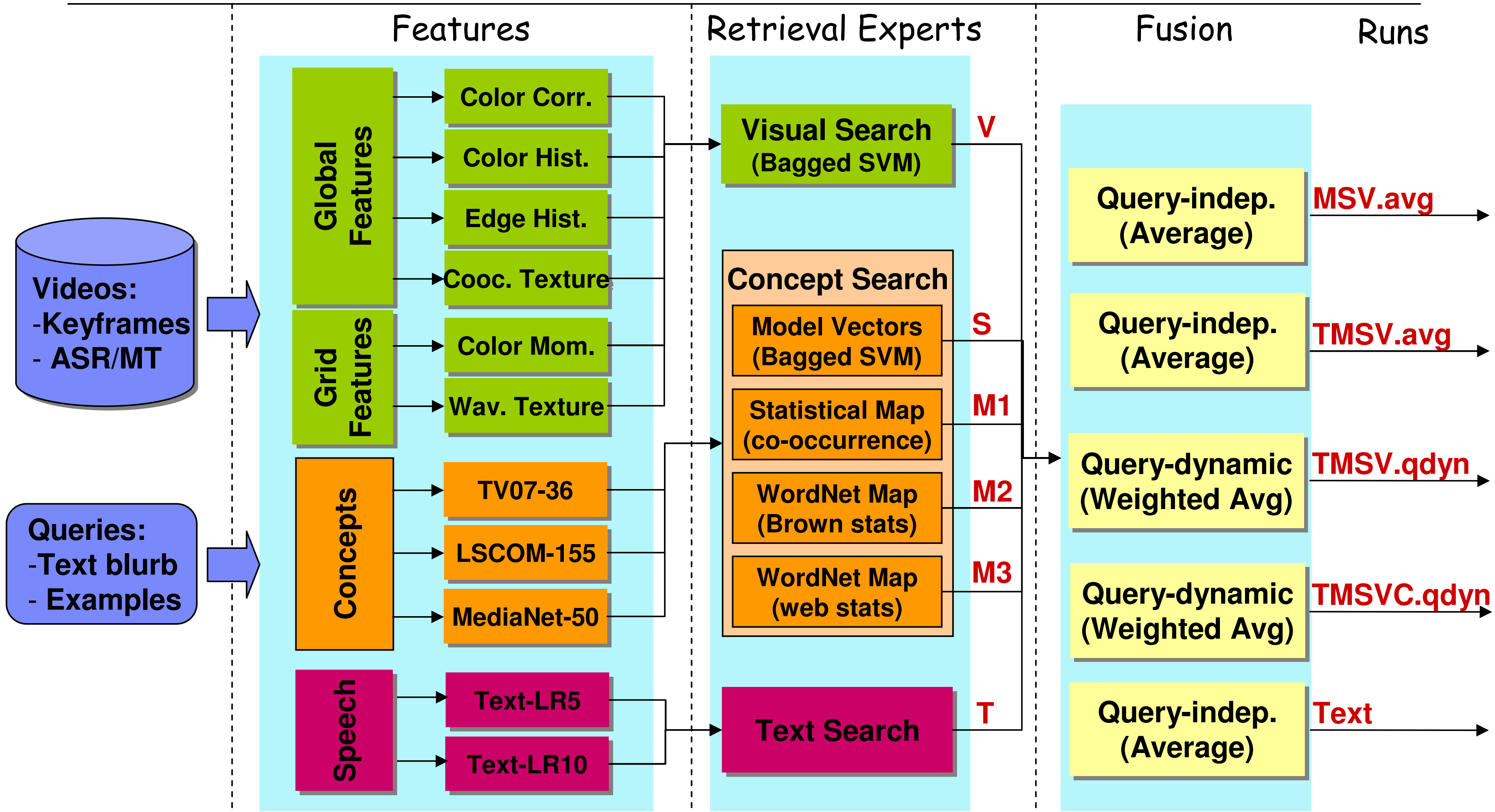
- **IBM Research**
  - Murray Campbell
  - Matthew Hill
  - Apostol (Paul) Natsev
  - Quoc-Bao Nguyen
  - Christian Penz
  - John R. Smith
  - Jelena Tesic
  - Lexing Xie
  - Rong Yan
- **UIUC**
  - Ming Liu
  - Xu Han
  - Xun Xu
  - Thomas Huang
- **Columbia Univ.**
  - Alexander Haubold
- **Carnegie Mellon Univ.**
  - Jun Yang

Part 1 of 2:
Automatic Search

Outline

- Overall System
- Summary
- Review of Baseline Retrieval Experts
- Performance Analysis
- Summary (Repeated)

System Overview



Summary: What Worked and What Didn't?

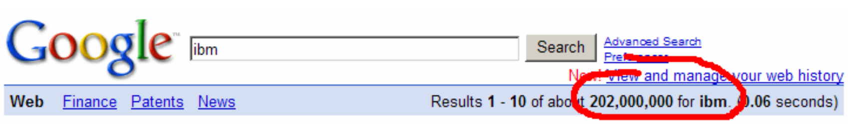
Baseline retrieval experts	Grade
Speech/text-based	✓
Concept-based (statistical mapping)	✗ ✗ ✗
Concept-based (WordNet mapping, Brown corpus stats)	✓
Concept-based (WordNet mapping, web stats)	✓ ✓
Concept-based (query examples modeling)	✓ ✓ ✓
Visual-based	✓ ✓ ✓
Experiments	
Improving query-to-concept mapping: web-based stats	✓ ✓
Leveraging external resources: type C runs	✓
Query-dynamic multimodal fusion	✗

Baseline Retrieval Experts: Review of Approaches

- Speech/Text-Based Retrieval
  – Auto-query expansion with JuruXML search engine (Volkmer et al., ICME'06)
- Visual-Based Retrieval
  – Bagged SVM modeling of query topic examples (Natsev et al., MM'05)
- Concept-Based Retrieval (G^2 Statistical Map)
  – Based on co-occurrence of ASR terms and concepts (Natsev et al., MM'07)
- Concept-Based Retrieval (WordNet Map, Brown Stats)
  – Based on JCN similarity, IC from Brown Corpus (Haubold et al., ICME'06)
- Concept-Based Retrieval (WordNet Map, Web Stats)
  – Based on JCN similarity, IC from WWW sample
- Concept-Based Retrieval (Content-Based Modeling)
  – Bagged SVM modeling of query topic examples (Tesic et al., CIVR'07)

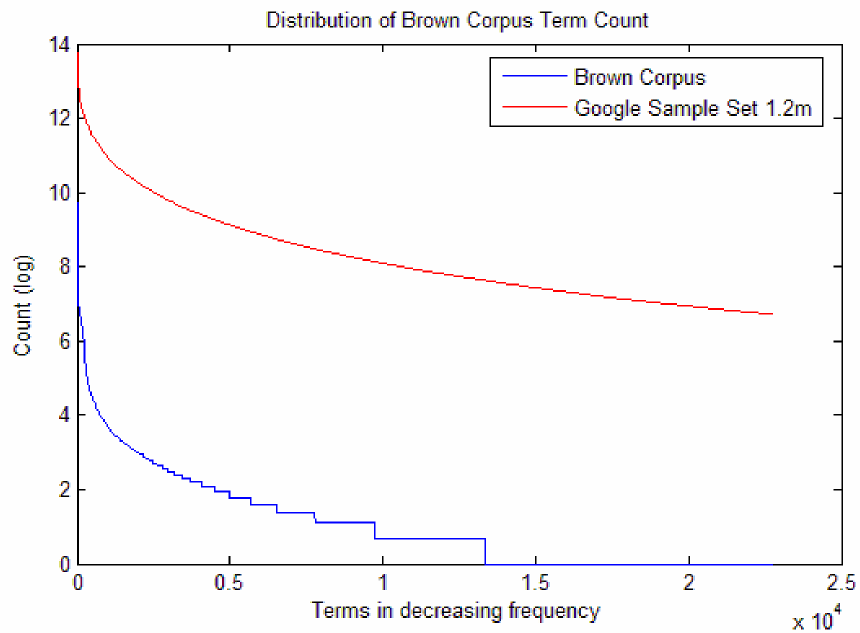
Improving Query-to-Concept Mapping

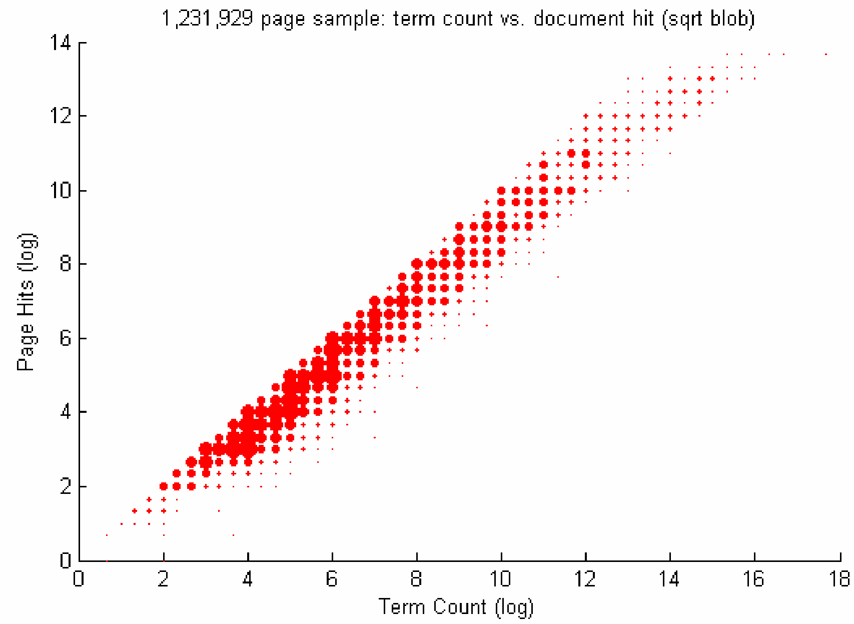
- WordNet similarity measures
  - Frequently used for query-to-concept mapping
  - Frequently based on Information Content (IC) to model term saliency
    - *Resnik*: evaluates information content (IC) of common root
    - *Jiang-Conrath*: evaluates IC of common root and ICs of terms
  - IC typically estimated from 1961 Brown Corpus
    - IC from Brown Corpus outdated by >40 years
    - 76% of words in WordNet not in Brown Corpus (so IC = 0)
- Idea: create approximation using WWW
  - Perform frequency analysis over large sample of web pages
  - Google page count as indicator of frequency



IC from Large-Scale Web Sampling

- Sample of web pages:
  - 1,231,929 documents (~1M)
  - 1,169,368,161 WordNet terms (~1B)
  - Distribution similar to Brown Corpus
  - Therefore: potentially useful as IC
- Google page count:
  - As a proxy to term frequency counts
  - Term frequency ≈ Doc. Frequency?
  - Experiments show linear relationship
  - Therefore: Potentially useful as IC

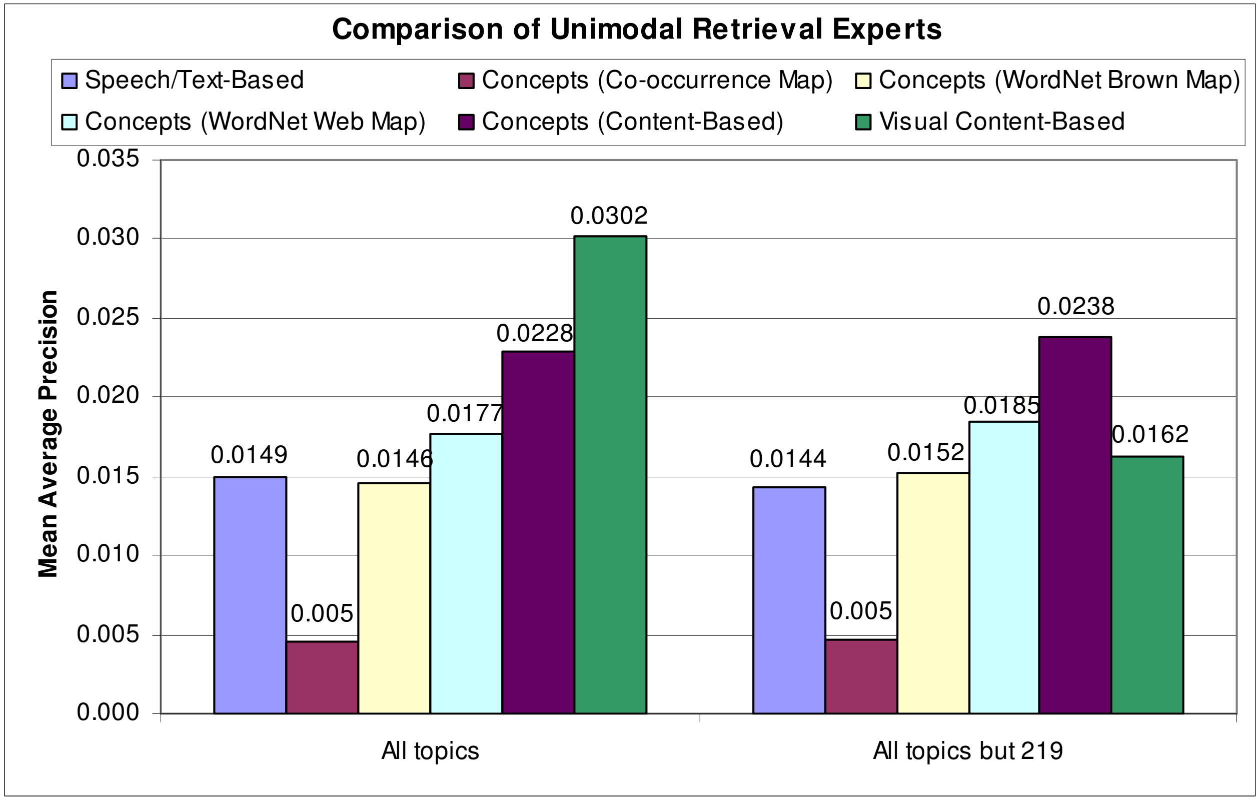




Outline

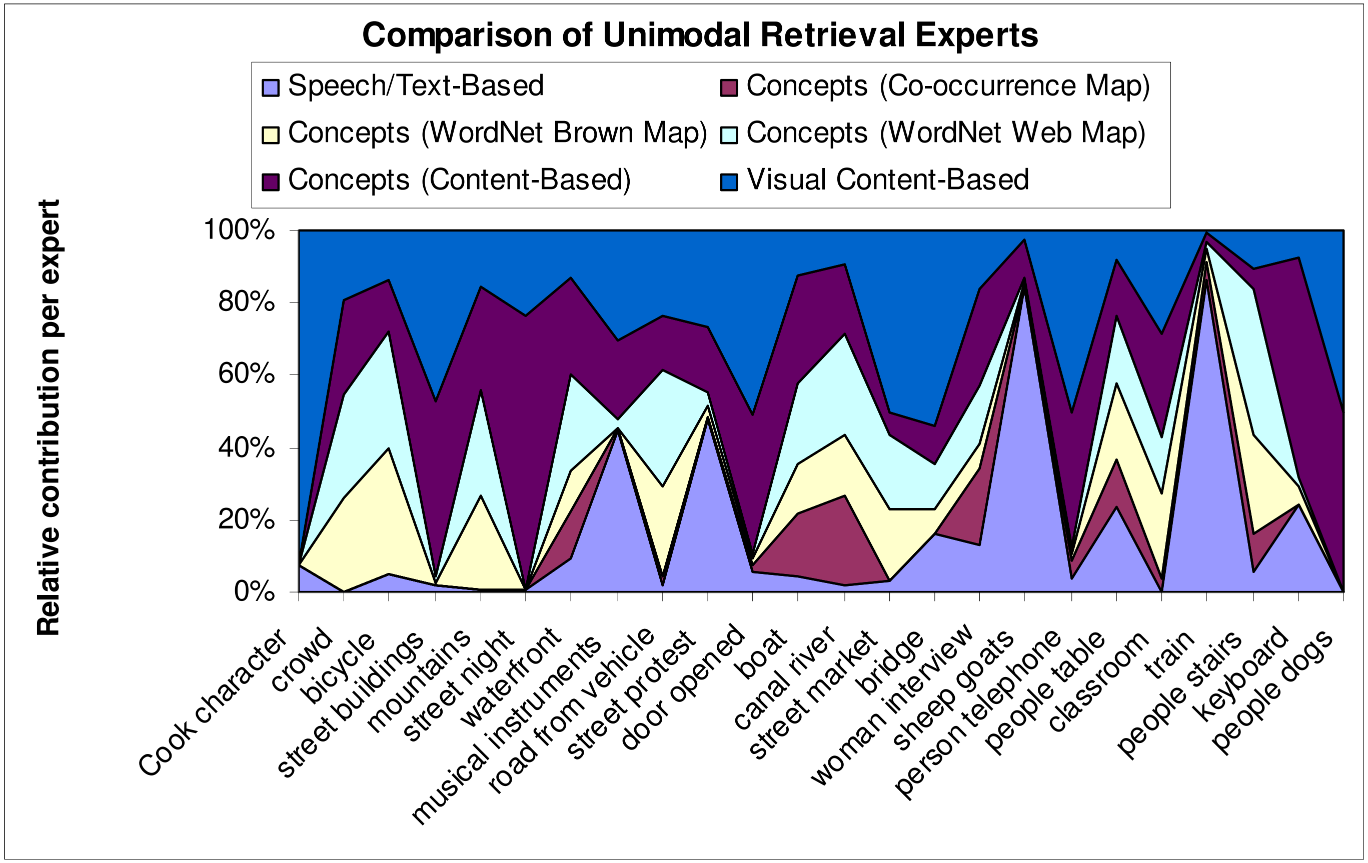
- Overall System
- Summary
- Review of Baseline Retrieval Experts
- ⇢ **Performance Analysis**
- Summary (Repeated)

Evaluation Results: Baseline Retrieval Experts



- Statistical concept-based run did not generalize
- Web-based IC led to 20% improvement in WordNet runs
- Concept-based runs performed better than speech/text-based runs
- Content-based runs performed best

Baseline Retrieval Experts: JAWS Analysis

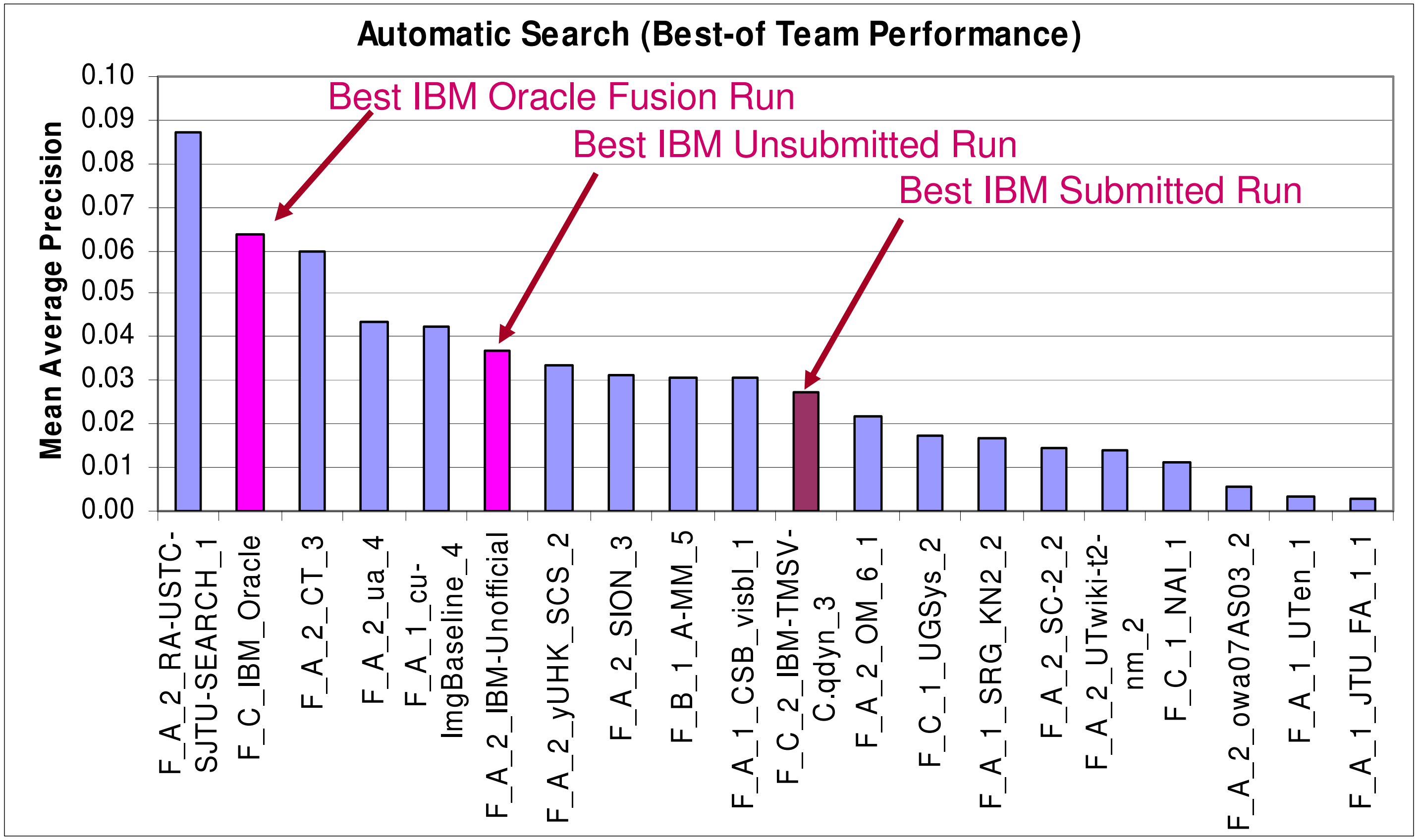


- For the adventurous:
  - Stacked chart showing contribution of each expert per topic

Summary of Submitted and Unsubmitted IBM Runs

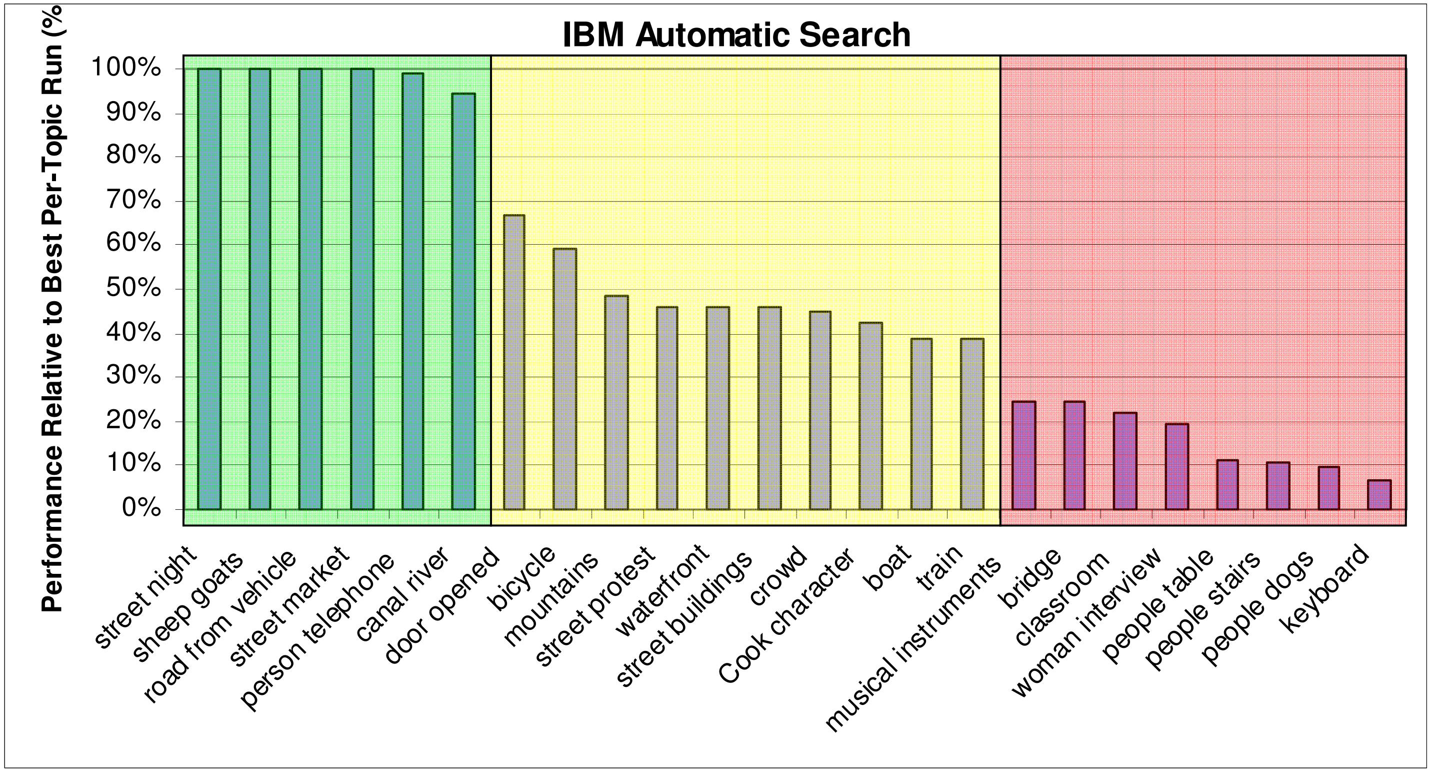
Description	Code	Run ID	Type	MAP
Text baseline	T	Text	A	0.0149
Concept baseline (stat)	M_S	-	A	0.0045
Concept baseline (WordNet, Brown stats)	M_B	-	A	0.0146
Concept baseline (WordNet, Web stats)	M_W	-	C	0.0177
Concept baseline (content-based, type A)	S	-	A	0.0228
Concept baseline (content-based, type C)	S_C	-	C	0.0249
Visual baseline	V	-	A	0.0302
Non-text baseline (M_S + S + V)	$M_S SV_{avg}$	MSV	A	0.0213
Non-text baseline (M_B + S + V)	$M_B SV_{avg}$	-	A	0.0301
Query-dynamic fusion, type A (T + M_S + S + V)	$TM_S SV_{qdyn}$	TMSV.qdyn	A	0.0210
Query-dynamic fusion, type C (T + M_{SW} + S_C + V)	$TM_{SW} S_C V_{qdyn}$	TMSV-C.qdyn	C	0.0272
Multimodal AVG fusion, type A (T + M_S + S + V)	$TM_S SV_{avg}$	TMSV	A	0.0231
Multimodal AVG fusion, type C (T + M_{SW} + S_C + V)	$TM_{SW} S_C V_{avg}$	-	C	0.0303
Multimodal AVG fusion, type C (T + M_W + S_C + V)	$TM_W S_C V_{avg}$	-	C	0.0372
Oracle fusion (pick best run for each topic)	Oracle	-	C	0.0638

Overall Performance Comparison



- Best run from each organization shown
- Submitted IBM runs in third tier, improve by dropping failed run
- IBM runs achieve highest AP scores on 5 of 24 topics

IBM Performance Relative to Best AP Per Topic



- Good (>90%): street night, street market, sheep/goat, road/vehicle, people telephone, canal/river
- So-so (40-60%): mountains, waterfront, boat, train, crowd, street protest, street buildings
- Bad (<25%): people table, people stairs, people dogs, people keyboard

Summary: What Worked and What Didn't?

Baseline retrieval experts	Grade
Speech/text-based	✓
Concept-based (statistical mapping)	✗ ✗ ✗
Concept-based (WordNet mapping, Brown corpus stats)	✓
Concept-based (WordNet mapping, web stats)	✓ ✓
Concept-based (query examples modeling)	✓ ✓ ✓
Visual-based	✓ ✓ ✓
Experiments	
Improving query-to-concept mapping: web-based stats	✓ ✓
Leveraging external resources: type C runs	✓
Query-dynamic multimodal fusion	✗

Part 2 of 2:
Interactive Search

Annotation-based Interactive Retrieval

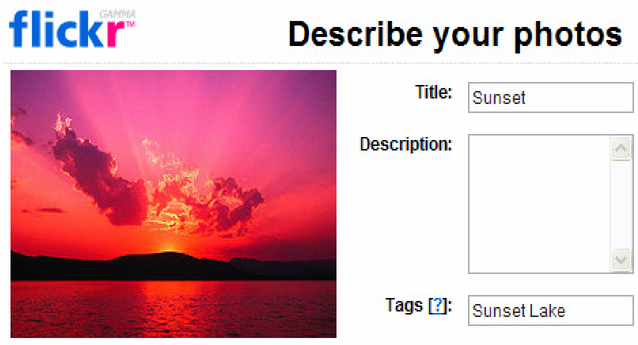
- Model interactive search as a video annotation task
  - Consider each query topic as a keyword, annotate video keyframes
  - Extend from CMU's Extreme Video Retrieval system [Hauptmann et al., MM'06]

- Hybrid annotation system
  - Minimize annotation time by leveraging two annotation interfaces
  - Tagging: Flickr, ESP game [von Ahn et al., CHI'04]
  - Browsing: IBM's EVA [Volker et al., MM'05], CMU's EVR [Hauptmann et al., MM'06]

- Formal analysis by modeling the interactive retrieval process
  - Tagging-based annotation time per shot
  - Browsing-based annotation time per shot

Manual Annotation (I) : Tagging

- Allow users to associate a single image at a time with one or more keywords (the most widely used manual annotation approaches)

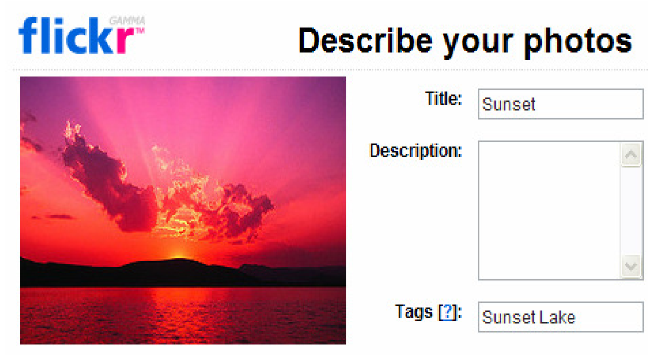


- Advantages
  - Freely choose arbitrary keywords to annotate
  - Only need to annotate relevant keywords
- Disadvantages
  - "Vocabulary mismatch" problem
  - Inefficient to design and type keywords

- Suitable for annotating rare keywords

Formal Time Model for Tagging

- Key factors for tagging time model
  - Number of keywords K for image l
  - Annotation time for k^{th} word t'_{fk}
  - Initial setup time for a new image t'_s
  - Noise term ε (zero-mean distribution)



- Annotation time for one image

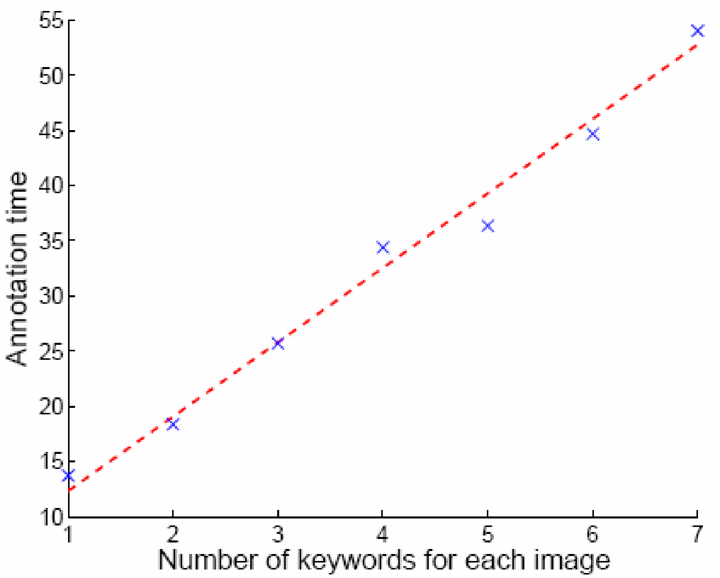
$$T = t'_{f1} + \ldots + t'_{fk} + t'_s + \varepsilon = \sum_k t'_{fk} + t'_s + \varepsilon$$

- Total expected annotation time for the entire collection containing L images
  - Assumption: the expected time to annotate the k^{th} word is constant t_f

$$E(T_{total}) = \sum_l \sum_{k_l} E\left(t'_{fk_l}\right) + E\left(t'_s\right) = \sum_l K_l t_f + t_s$$

Validation of Tagging Time Model

- User study on TRECVID'05 development data
  - A user to manually tag 100 images using 303 keywords
  - If the model is correct, a linear fit should be found in the results
  - The annotation results fit the model very well



Manual Annotation (II) : Browsing

- Allow users to associate multiple images with a single word at the same time

- Advantages
  - Efficient to annotate each pair of images and words
  - No “vocabulary mismatch”

- Disadvantages
  - Need to spend time on judging both relevant and irrelevant pairs
  - Start with controlled vocabulary
  - Annotate one keyword at a time

- Suitable for annotating frequent keywords

Formal Time Model for Browsing

- Key factors for browsing time model
  - Number of relevant images L_k for a word k
  - Annotation time for a relevant image t'_p
  - Annotation time for an irrelevant image t'_n
  - Noise term ε (zero-mean distribution)

- Annotation time for all images w.r.t. a keyword

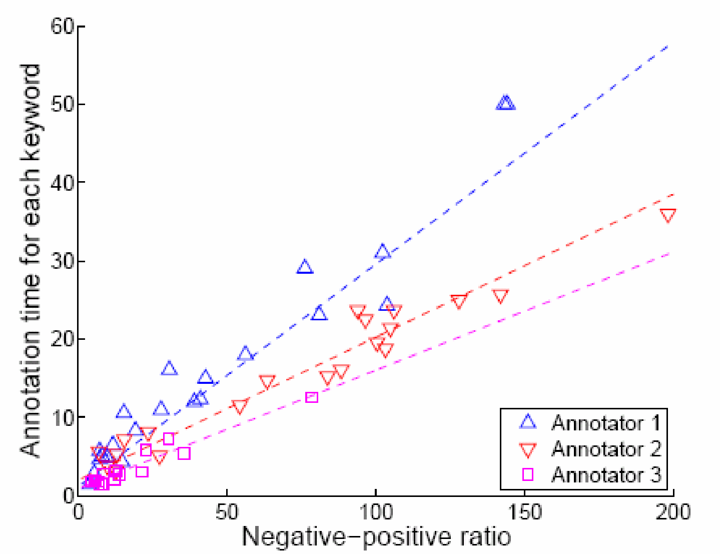
$$T = \sum_{l=1}^{L_k} t'_{pl} + \sum_{l=1}^{L-L_k} t'_{nl} + \varepsilon$$

- Total expected annotation time for the entire collection containing L images
  - Assumption: the expected time to annotate a relevant (irrelevant) image is constant

$$E(T_{total}) = \sum_k \left(\sum_{l_k} E\left(t'_{pl_k}\right) + \sum_{l_k} E\left(t'_{nl_k}\right) \right) = \sum_k \left(L_k t_p + \left(L - L_k\right) t_n \right)$$

Validation of Browsing Time Model

- User study on TRECVID'05 development data
  - Three users to manually browse images in 15 minutes (for 25 keywords)
  - If the model is correct, a linear fit should be found in the results
  - The annotation results fit the model for all users

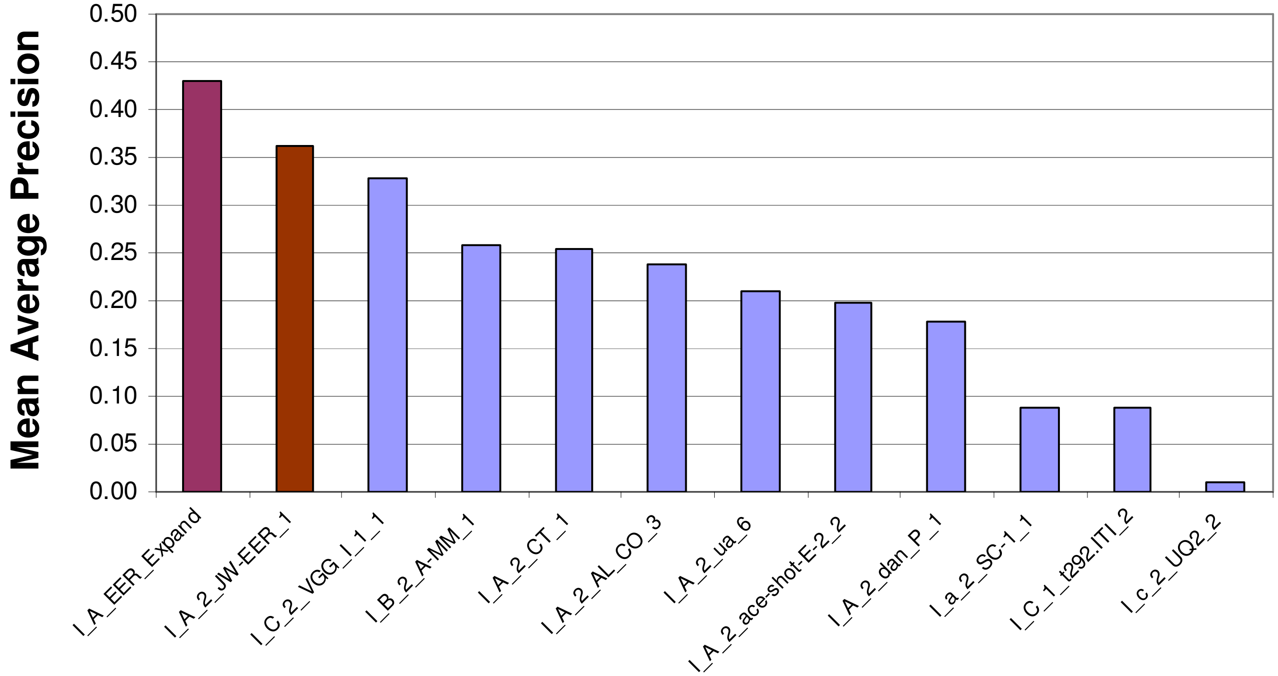


Video Retrieval as Hybrid Annotation

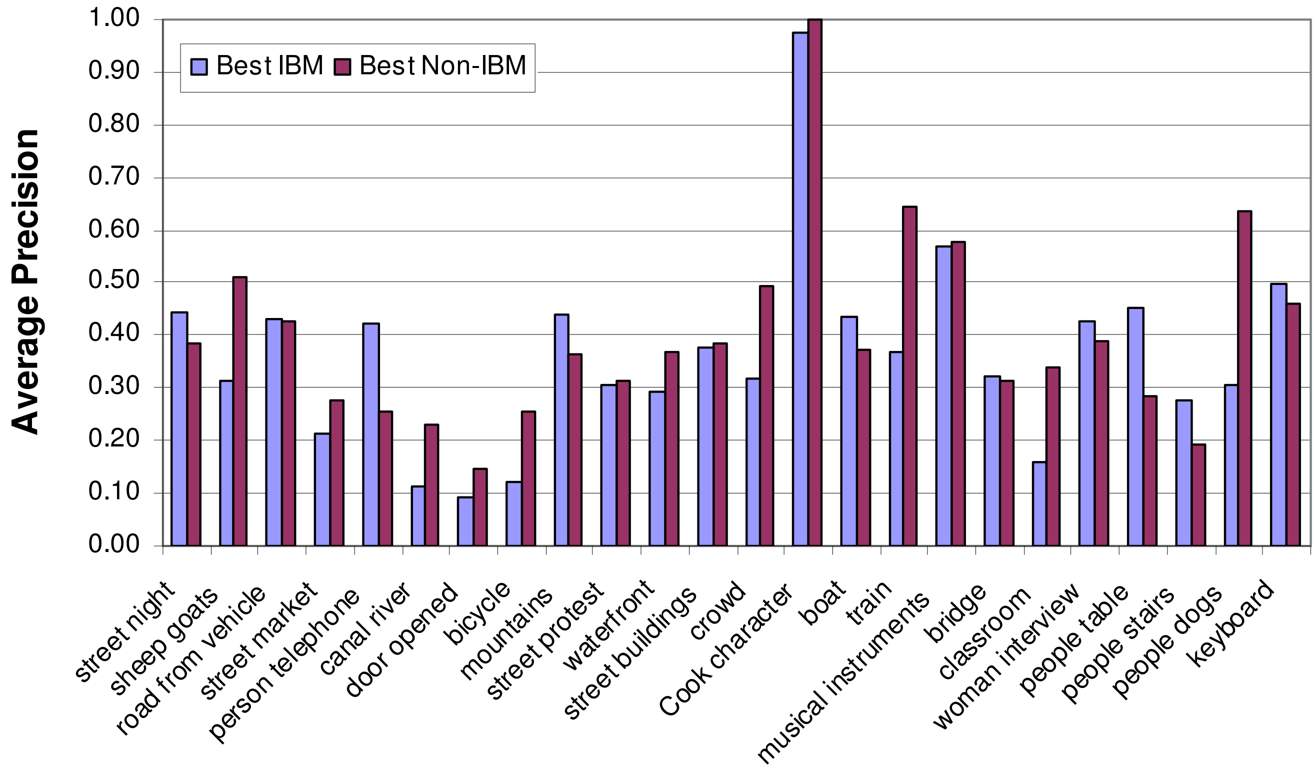
- Jointly annotates all topics at the same time
- Switches between tagging and browsing annotation interfaces in order to minimize the overall annotation time
  - Formally model the annotation time as functions of word frequency, time per word, and annotation interfaces
  - Online machine learning to select images, topics and interfaces based on the annotation models
- More details will be released in the final notebook paper
  - See related analysis in R. Yan et al. [ACM MM 2007 Workshop on Many Faces of Multimedia]

TRECVID-2007 Performance Analysis

- The proposed approach allows users to annotate 60% of the image-topic pairs, as compared with ~10% allowed by simple browsing
- Balance between tagging & browsing: 1529 retrieved shots from tagging, 797 retrieved shots from browsing
- Simple temporal expansion can improve MAP from 0.35 to 0.43



TRECVID-2007 Per-Query Performance Analysis



- Good: “telephone”, “interview”, “mountains”
- Not as good: “canal”, “bicycle”, “classroom”, “dog”
- Highest AP scores on 10 out of 24 topics

Potential Improvement

- Search beyond **keyframe** (look at the I-frame)

Finding Sheep/Goats

- Search beyond **I-frame** (leverage text info or temporal context)

Finding Sheep/Goats

- Better online learning algorithms to search for more shots in the same amount of time (include temporal information)