informedia
digital video understanding
SEARCH summarize visualize retrieve

Machine Learning, Pattern Recognition, Cross-modal Analysis and Fusion

Alex Hauptmann
School of Computer Science
Carnegie Mellon University,
Pittsburgh, PA, USA
alex@cs.cmu.edu
http://www.cs.cmu.edu/~alex

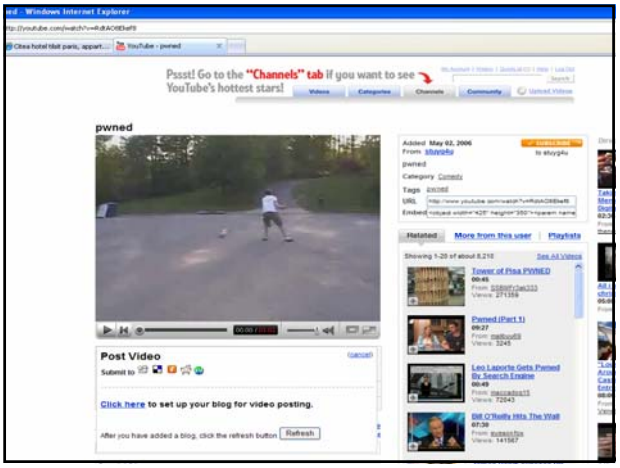
Carnegie Mellon

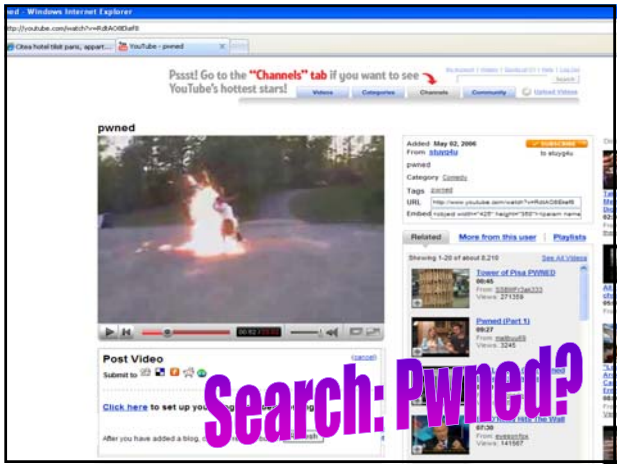
Outline

Part Two

- Bridging the Semantic Gap across Modalities
  - A Large Scale Ontology for Multimedia (LSCOM)
  - Retrieval Experiments with LSCOM
  - Active Learning of Semantic Concepts
- YouTube and Challenges for the Future

Carnegie Mellon





TRECVID Benchmark Evaluation

Goal is to promote progress in content-based retrieval from digital video via open, metrics-based evaluation

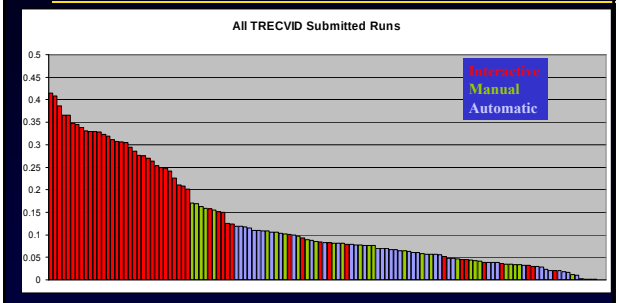
- Laboratory style evaluation, modeled on real world
- Emphasis on good science
- Evaluation is driven by participants (and feasibility)
  - More than 60 participating groups in 2007
- Collection is fixed, available in the spring

  50% data used for development, 50% for testing
- ~24 Test queries available in July, 1 month to submission

Details at http://www-nlpir.nist.gov/projects/trecvid

Carnegie Mellon

TRECVID 2005 System Comparisons

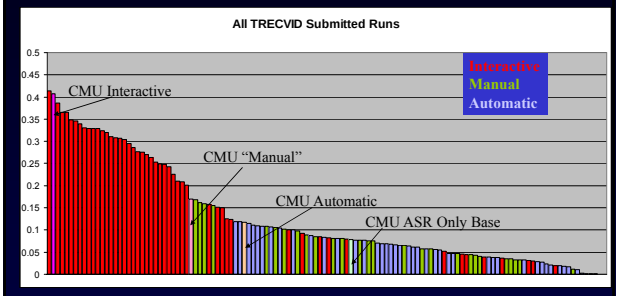


Differences between the top non-interactive systems are not significant

Accuracy for non-interactive systems is consistently **LOW**

Carnegie Mellon

TRECVID 2005 System Comparisons



Differences between the top non-interactive systems are not significant

Accuracy for non-interactive systems is consistently **LOW**

Carnegie Mellon

Research in Semantic Concepts

"High-Level" or Semantic Features

Faces, cars, outdoors, people, buildings, vegetation, rivers, animals, …

- Can be detected automatically
  from color, texture, shape, style
- Intense competition for good classifiers
  (TRECVID)
- Useful for retrieval and summarization

Carnegie Mellon

LSCOM Large Scale Ontology for Multimedia

Problem:

- Users and analysts require richly annotated video content for search and retrieval
- We don't know how to translate video content into words
- Manual annotation is prohibitively expensive and slow

Solution:

- Find a restricted (controlled) concept vocabulary which can be used to (automatically) describe broadcast news video content
  - Start with 1000 concepts grouped into a taxonomy/ontology
  - Evaluate if these concepts are useful for retrieval
  - Test if they can be automatically detected
  - Iterate

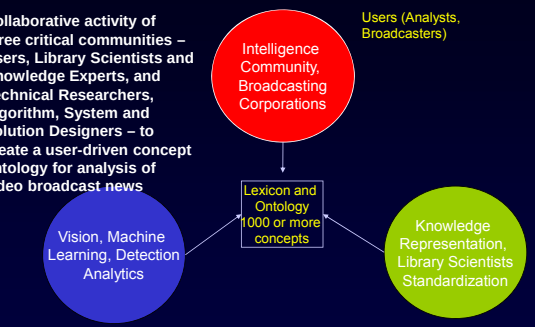
Impact:

- Allow useful classification of multilingual broadcast video
- Provide an extensible framework and procedures for video analysis, beyond the 1000 concepts

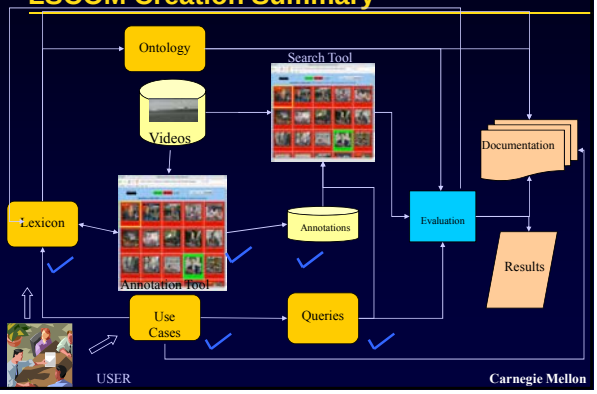
Carnegie Mellon

Central Idea

- **Collaborative activity of three critical communities – Users, Library Scientists and Knowledge Experts, and Technical Researchers, Algorithm, System and Solution Designers – to create a user-driven concept ontology for analysis of video broadcast news**



LSCOM Creation Summary



Carnegie Mellon

Multimedia Controlled Vocabularies

- MPEG-7 Multimedia Description Schemes standardizes classification schemes comprising 754 controlled terms
- TV-Anytime standardizes 954 terms for broadcast TV content
- Escort 2.4 defines 115 terms
- Thesaurus of Graphical Materials (TGM-I) defines index terms for cataloging graphical (image and video) content
- SMPTE Metadata Registry from Society of Motion Picture and Television Engineers has a metadata dictionary defining a registry of metadata element descriptions for association with essence or other metadata. (Full explanation in SMPTE 335M)
- IPTC Core Metadata, NewsML, SportsML and ProgramGuide ML from International Press Telecommunications Council. The versatile News Markup Language for global news exchange. NewsML 1 is designed to provide a media-independent structural framework for multimedia news.

Carnegie Mellon

LSCOM Lexicon Design

DESIGN
Starter Set
Consultation
Filtering
Sample Verification
Lexicon

More than 30 Media Analytics Experts, 10 User Community Experts and 6 Knowledge Representation Experts met twice

More than 10,000 concepts
TGM, Time Life, TV Anytime, Comstock, WordNet

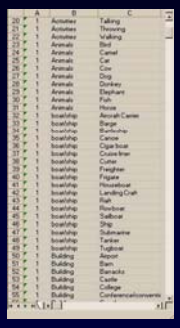
More than 600 concepts from media companies, intelligence analysts

Filtered down to 834 concepts (so far)
based on Usability, Feasibility and Observability

Manual annotation over sample video corpus led to
Annotation of 449 unique concepts (so far) based on availability of concept in corpus and inter-annotator agreement

Carnegie Mellon

Sample Lexicon



Design Criteria
- Utility
- Feasibility
- Observability

Additional Annotation Criteria
- Availability in the Corpus
- Inter-annotator Agreement

Carnegie Mellon

Annotated Data

- TRECVID 2005 development set
  - 80 hours of broadcast news
    - 11 programs
    - 6 channels
    - 3 languages
  - 61901 shots
  - Each concept (449, total) labeled over each shot.

Carnegie Mellon

Connecting Concept to an Ontology

- Cyc – more than 20 years building an ontology of everything
- Specialized and more useful for specific domains

Carnegie Mellon

Mapping the LSCOM Taxonomy into CYC

HL Formula :

Mt : LSCOMObjectAndSituationOntologyMappingMt
(synonymousExternalConcept Place LSCOMObjectAndSituationOntology "locations")

Mt : LSCOMObjectAndSituationOntologyMappingMt
(synonymousExternalConcept Island LSCOMObjectAndSituationOntology "island")

Then map the graphical structure.

Mt : (ContextOfPCWFn LSCOMObjectAndSituationOntology)
(broaderThan Island Place)

Carnegie Mellon

Cyc can use the mappings to reason about the LSCOM taxonomy

EL Query :
(and
(synonymousExternalConcept ?X LSCOMObjectAndSituationOntology ?S)
(sourceNodeInSystem ?X LSCOMObjectAndSituationOntology))

Status : Suspended, Exhaust

Hide **Answers** *(11 new)* **Actions :** [Save Answers] [Graph Answers] [

Answer	?X	?S
[Explain #10]	HumanScaleObject	"Objects"
[Explain #9]	Person	"People"
[Explain #8]	Roadway	"Road"
[Explain #7]	AnimateActivity	"activities"
[Explain #6]	Barrier-ConstructionArtifact	"Barrier"
[Explain #5]	Event	"Events"
[Explain #4]	AirTransportationDevice	"flying-object"
[Explain #3]	Place	"locations"
[Explain #2]	PlayingOf-SportsEvent	"sports"
[Explain #1]	Building	"Building"
[Explain #0]	SewerSystem	"Sewage infrastructure"

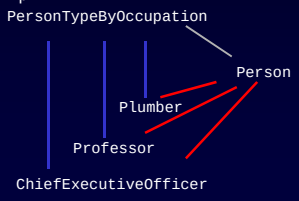
Results of a query for the top (source) nodes in the LSCOM taxonomy, as authored by workshop participants.

Carnegie Mellon

Cyc can Suggest Extensions/Improvements

Use of Faceting Collections

- Higher-order collections that facet a class into sub-classes whose instances share a property
- Example:



Carnegie Mellon

Final Statistics of the Ontology

Size: 784 → 2638

Annotated: 450

Distribution through LSCOM.org

Carnegie Mellon

What are the Limits of Semantic Concepts

- How much can semantic concepts improve video search? What is the upper bound?
- How many do we need for good retrieval?
- Are there enough concepts out there?
- Can we find them?

Carnegie Mellon

Can *Semantic Concepts* provide significant boost to IR?

- Low level visual features are not sufficient to understand an image or video clip ("Semantic Gap")
  - Low-level: Texture, color, shape, interest points, motion, audio (SFFT, MelCep, Zero crossing, …)
- Describe video through intermediate ***semantic*** concepts
  - Face, car, outdoors, boat, building, clouds, sky, water, …
- Semantic concepts can be learned automatically
- Semantic concepts are useful for retrieval and summarization

Carnegie Mellon

Why are Semantic Concepts Important?

- What if we could detect a lot of concepts?
- Speech recognition analogy
  - 100 words → 1000 words → 20,000 words → 64,000 words
- LSCOM – A Large Scale Ontology for Multimedia
  - 2 year workshop to define and annotate 1000 concepts
  - Defined 850 concepts
  - Extended via ontology to ~2400 concepts,
  - Annotated 450 concepts on 70 hours of TV news
  - Available at www.LSCOM.org

Carnegie Mellon

Building a Semantic Concept Detector

- Get training data
  - Manual labels over a representative, large data set
- Extract feature vectors
  - Color, texture, shape, motion, feature points, text, audio
  - Very high dimensional, continuous features
- Run machine learning algorithm
  - SVM, AdaBoost, Decision Trees, Graphic Models, etc.

Presto!

But there is a lot of detail to be considered…

- Details matter: feature selection, parameter tuning, model combination

Carnegie Mellon

Video Data: TrecVid 2006 Development Corpus

- About 70 hours of multilingual broadcast news from October – November 2004
- English, Chinese, Arabic
- 61901 shots as the units of analysis
- Truth annotations for about 450 LSCOM concepts

- Evaluation by NIST on a separate test corpus for detection of a subset of the semantic concepts

Carnegie Mellon

39 Semantic Concepts (LSCOM-Lite)

1	Sports	20	Person
2	Entertainment	21	Government-Leader
3	Weather	22	Corporate-Leader
4	Court	23	Police-Security
5	Office	24	Military
6	Meeting	25	Prisoner
7	Studio	26	Animal
8	Outdoor	27	Computer-TV
9	Building	28	Flag-US
10	Desert	29	Airplane
11	Vegetation	30	Car
12	Mountain	31	Bus
13	Road	32	Truck
14	Sky	33	Boat-Ship
15	Snow	34	Walking-Running
16	Urban	35	People-Marching
17	Waterfront	36	Explosion-Fire
18	Crowd	37	Natural-Disaster
19	Face	38	Maps 39 Charts

Carnegie Mellon

Annotated Concept Sets

- Trecvid 2006 development data
  - ~70hours English, Arabic, Chinese News
  - 62000 shots

3 Annotated Concept Sets:
- LSCOM Lite
  - 39 concepts

- Media Mill
  - 75 concepts that overlap with LSCOM

- LSCOM
  - 300 concepts
  - Minimal frequency cutoff

Carnegie Mellon

A Speculative Scenario

What if we could detect a lot of concepts?

Best Case:
- Perfect concept detection (Oracle)
- Perfect concept combination (Oracle)

How well can you retrieve relevant shots (documents)

Carnegie Mellon

Computing the Optimal Weights

$$LLB = \max_{\lambda_i} AP\left(\sum_{i=1}^{k} \lambda_i f_i(D,Q)\right)$$

This is an upper bound under perfect conditions

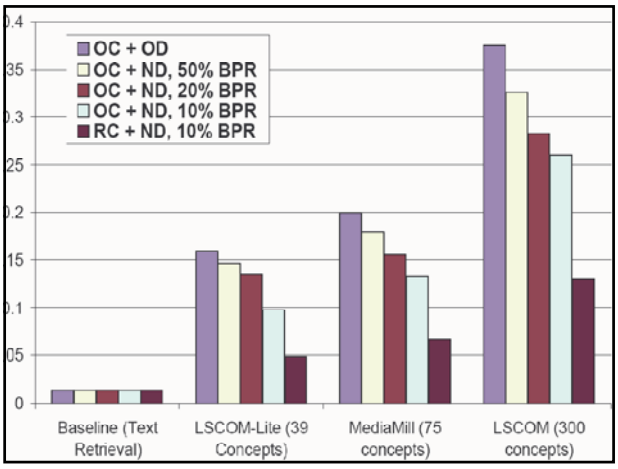
Carnegie Mellon

Assumptions for Realistic Evaluation

- Noisy Detection - Assume MAP of 0.5, 0.2 or 0.1
  - TRECVID 2007 semantic concept detection results just under 0.2
- Approximate through precision/recall breakeven point at 50%, 20%, 10%

- Realistic combination (RC)
  - achieves only 50% of perfect (Oracle) combination
  - Based on prior TRECVID experiments

Carnegie Mellon



Extrapolation to thousands of concepts

- How many concepts do we need?
- Exponential function to fit curve
  - Only 3 points

Assumption:

- Things get harder as you add more concepts
  - Proportional to the difference between the current MAP and the upper limit of 1
    I.e. the higher the current MAP, the less benefit a new concept offers

Extrapolation Assumptions

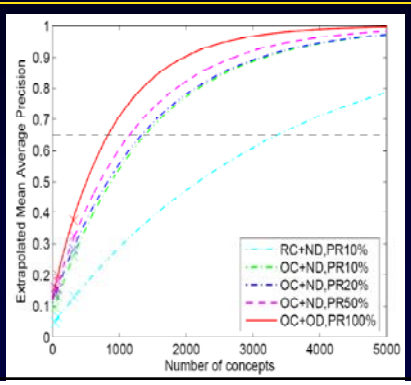
$$\frac{dm}{dx} \propto (1-m)$$

where m is the value of MAP, x is the number of concepts, and the boundary condition is m(infinity) = 1.

This yields:

$$m(x) = 1 - \exp(ax+b)$$

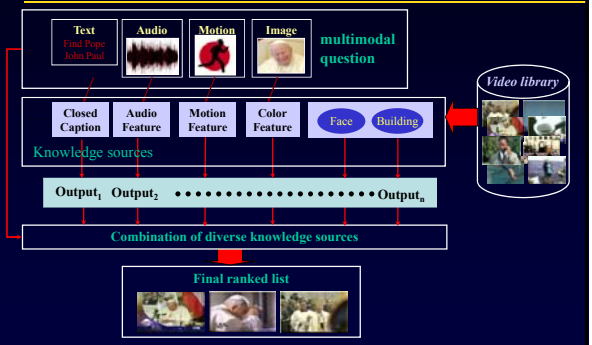
Extrapolation Results



Conjecture: ~4000 concepts are enough

[CIVR2007]

Combining Concept Detectors for Retrieval



Probabilistic Model for Video Retrieval

- What is the probability *this* shot (document) is an answer for *this* query?

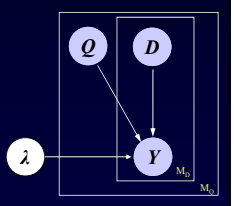
Estimate P(Y=1|D,Q)

- Basic model: logistic regression for multimedia IR
  - For every document *D* and query *Q*,

$$P(Y=1 \mid D,Q) = \sigma\left(\sum_{i=0}^{N} \lambda_i P(S_i \mid D,Q)\right)$$

$$where \quad \sigma(x) = \left(1+e^{-x}\right)^{-1}$$

  - Learn the combination weights λ_i from past retrieval results



"Ranking" Logistic Regression

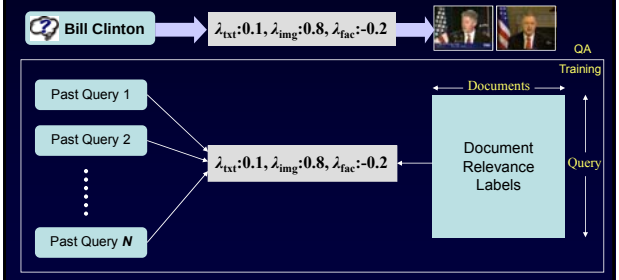
- Problem: logistic regression is designed for classification
  - Simplify the QA problem to binary classification
  - Several drawbacks, e.g. bias to irrelevant documents
- Our model: "ranking" logistic regression

$$\max_{\lambda} \sum_{q \in Q} \sum_{d_1 \in D^+} \sum_{d_2 \in D^-} \log \sigma\left(\sum_{i=0}^{N} \lambda_i \left[P(S_i \mid d_1, q) - P(S_i \mid d_2, q)\right]\right)$$

  - Maximize the prediction gaps between positive/negative examples
  - This is a lower bound of the *average precision*
- Approximate inference
  - Reduce from $O(n^2)$ to the same complexity as logistic regression

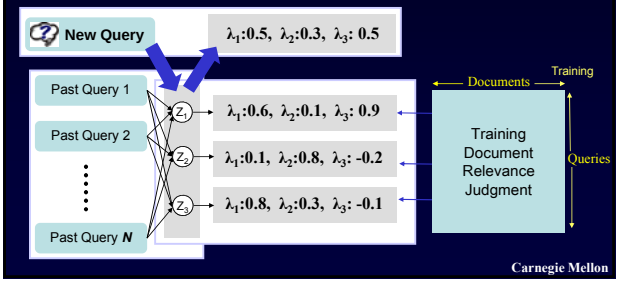
Query Independent Combination

- Learn a single combination function from training queries
- Uniform combination function for all other queries



Probabilistic Latent Query Analysis (pLQA) [Yan06]

- Each query can be described with a mixture of latent query classes
- Each query class is associated with a set of combination weights
- Query description is used to predict to which query class it belongs

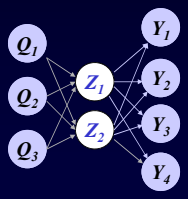


Basic pLQA (BpLQA)

- BpLQA: learn the query classes directly from training data
  - Introduce new latent class variables Z in the logistic regression model
  - Query classification and combination optimization in a unified framework

$$P(y \mid D, Q) = \sum_{z} P(z \mid Q; \mu) \cdot \sigma\left(y \sum_{i=0}^{N} \lambda_{zi} f_i(D, Q)\right)$$

- Advantages over query-class combination
  - No manual definition on query classes
  - Allow mixture of query classes for one query
  - Able to find optimal number of query classes using model selection criteria, e.g. AIC, BIC



Adaptive pLQA (ApLQA) Extension to Unseen Queries

- **Problem:** P(z|Q;μ) is associated with each training query
- **Goal:** extend basic pLQA to handle unseen queries outside the training collection
- ApLQA: represent P(z|Q;μ) as a soft-max function over a vector of predefined query features q_l
  - *e.g.*, whether the text query contain a specific person name
  - *e.g.*, how many objects are mentioned in the query
  - *e.g.*, the mean of text retrieval scores

$$P(y \mid D, Q) \propto \sum_{z} \exp\left(\sum_{l=0}^{L} \mu_{zl} q_l\right) \cdot \sigma\left(y \sum_{i=0}^{N} \lambda_{zi} f_i(D, Q)\right)$$

Parameter Estimation for ApLQA

- Bayesian information criterion (BIC) to select the optimal number of query classes
- Maximum likelihood estimation using the EM algorithm
  - Initialize combination weights from automatically selected queries
  - **E-step:** compute the expectation of latent query type *Z*

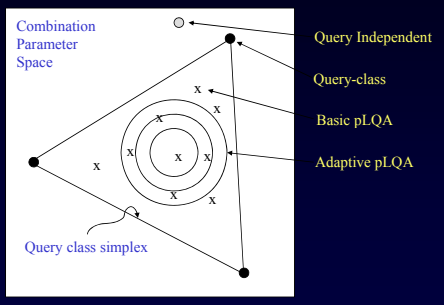
$$P(z_k \mid Q_t, D_j) = \frac{\exp\left(\sum_{l=0}^{L} \mu_{kl} q_l\right) \cdot \sigma\left(y_{tj} \sum_{i=1}^{N} \lambda_{ki} f_i(Q_t, D_j)\right)}{\sum_{k=1}^{K} \exp\left(\sum_{l=0}^{L} \mu_{kl} q_l\right) \cdot \sigma\left(y_{tj} \sum_{i=1}^{N} \lambda_{ki} f_i(Q_t, D_j)\right)}$$

  - **M-step:** estimate the parameters using posterior probabilities of *Z*

$$\lambda_{zi} = \arg\max_{\lambda} \sum_{t=1}^{M_Q} \sum_{j=1}^{M_D} P(z \mid Q_t, D_j) \log\left[\sigma\left(y_{tj} \sum_{i=1}^{N} \lambda_{zi} f_i(Q_t, D_j)\right)\right]$$

$$\mu_{zl} = \arg\max_{zl} \sum_{t} \left(\sum_{j} P(z_k \mid Q_t, D_j)\right) \log\left[\frac{1}{Z_l} \exp\left(\sum_{l} \mu_{zl} q_{tl}\right)\right]$$

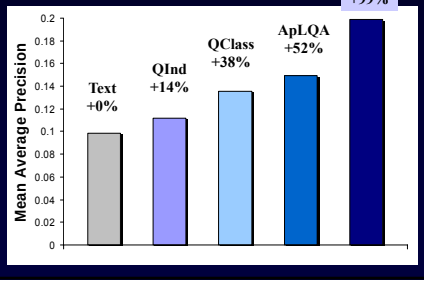
Geometric Interpretation



2 combination parameters and 3 query classes

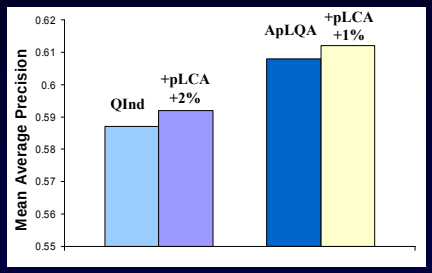
Video Retrieval Performance

- Average retrieval performance on TREC'02-'05 test sets
- ApLQA can achieve a higher performance without manually defining the query classes



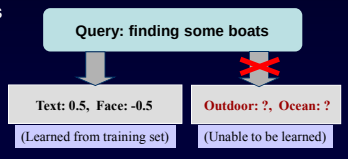
Meta-Search Performance

- Performance of pLQA on TREC-8 testing sets
  - pLCA on 2 retrieval algorithms w. 5 additional ranking features



Limitations of pLQA

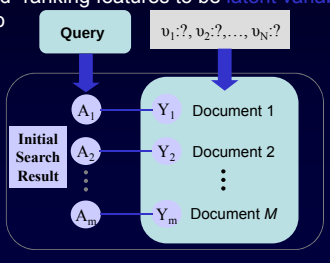
- Due to a limited amount of training data, the weights of many ranking features are simply set to zero
- More than 80% of the ranking features are ignored in TREC'05
- However, these features are useful for some specific queries



Let's try relevance feedback (local content analysis)

Probabilistic Local Context Analysis (pLCA) [Yan06]

- **Goal:** automatically leverage useful features for current query
- **Method:** assume combination parameters υ of "un-learned" ranking features to be latent variables instead of zero



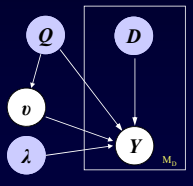
Model and Parameter Estimation

- Jointly optimize weights of query-specific sources ***u*** and answer relevance ***Y*** in initial retrieved shots

$$\max_{y,v} P(v \mid Q) \cdot \prod_{j=1}^{M_D} \sigma\left(\sum_{i=0}^{N} y_j \lambda_i P(S_i \mid D_j, Q) + \sum_{l=1}^{M} y_j v_l P(S_l^* \mid D_j, Q) \right)$$

(Prior) (Query Independent) (Query Specific)

- Iteratively maximize ***Y*** and ***u***
  - ***Y***: top-ranked shots as positive and others as negative
  - ***u***: feature selection + regularized logistic regression

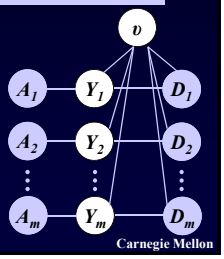


pLCA: Undirected Model and Parameter Estimation

- Compute the posterior probability of document relevance ***Y*** given initial results A based on an undirected graphical model

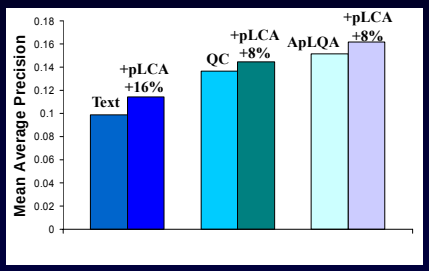
$$P(\vec{y}\mid\vec{a},\vec{D},Q)=\frac{1}{Z}\int_{v}\prod_{i}P(v_i\mid Q;v_i^0)\cdot\prod_{j=1}^{M_D}\exp\left(y_j a_j+y_j\sum_i v_i f_i(D_j,Q)\right)dv$$

- Variational inference
  - Approximate the posterior distribution of **p(y,v|a)** by a family of variational distributions **q(y,v)** where y and v are independent
  - Iteratively maximize a variational lower bound of the log-likelihood function until it converges



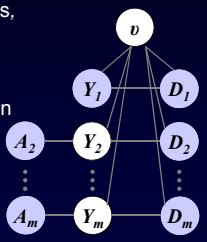
Video Retrieval Performance

- Retrieval performance of pLCA on TREC'03-'05
  - pLCA on 3 retrieval algorithms w. ~50 additional ranking features



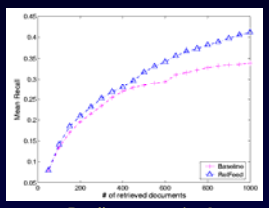
pLCA with Relevance Feedback

- Users could provide their own relevance judgments in an interactive retrieval interface.
- Given these relevance feedbacks, we can use a similar variational inference technique to update parameters except that some relevance variables $\mathbf{Y_i}$ are fixed in this case.

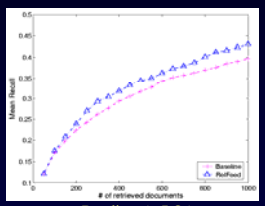


Results of pLCA Relevance Feedback

- Performance in terms of average recall
- Update combination parameters every 50 shots



Baseline: **text retrieval**



Baseline: **ApLQA**

Extreme Video Retrieval (Interactive)

- Maximizing the Synergy between Man and Machine

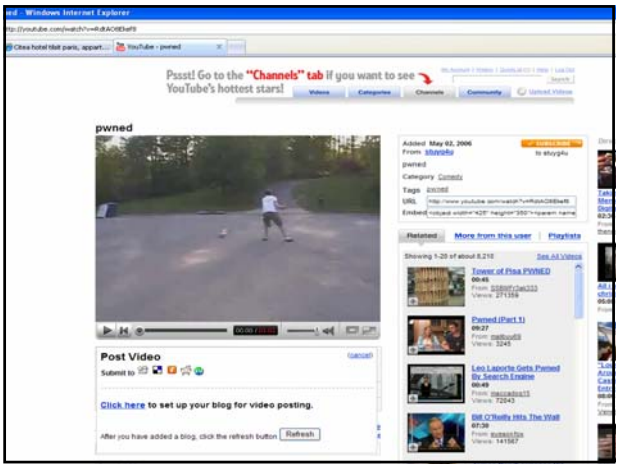
- Maximizes information transfer: System → Human
  - Up to 10 keyframes/second
  - 1 - 9 images per page (dynamically adjustable)
  - Presentation intervals
    - Automatic: System presents next page after X msecs (adjustable)
    - Manual: Press key for next page
  - Click where relevant shot is seen

- Relevance feedback based on user clicks
- Automatic retrieval baseline for ranking order

- Requires full concentration: 3000-5000 keyframes in 15 minutes

- Second best interactive system in TrecVid 2005
- Fourth best in TrecVid 2006
- Best system in Video Olympics 2007 (NUS/CAS)

Challenges for the Future



YouTube

- Annotations as user tags and descriptions
- Not very accurate, but abundantly available
- Exploit other information
  - Author
  - User
  - Comment
  - Video Responses
- Google scale

Near Term Research (5 -10 years)

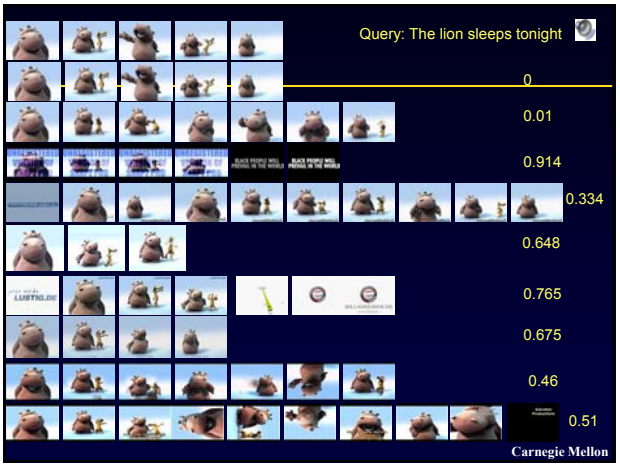
- Leverage the web
  - 'Public' data
  - Social tagging
  - Games
- Bridging the Semantic Gap (perhaps using Concepts)
  - Better retrieval
- Human computer interaction
  - better AV search interfaces
- Combining world knowledge with AV analysis for search and summaries
  - Ontologies?
- "Event Search" (R. Jain)
- Privacy, digital rights
- Data Integrity
- Data scale increased to 'Google-scale'

Data Integrity of Web Content

Near-duplicate and tamper detection
for security, copyright, ease of browsing

Types of Near-Duplicates and Tampering:

- Reformatting (frames rate, frame size,
- Mixing old video with new audio (or vice versa)
- Image variations: color, lighting, cropping of frame
- Adding/Deleting frames, mostly at the head and tail
- Insertion of new pieces
- Re-editing for different versions: 1'08", 2'50"
- Overlay text and logos
- Recomposition of image components

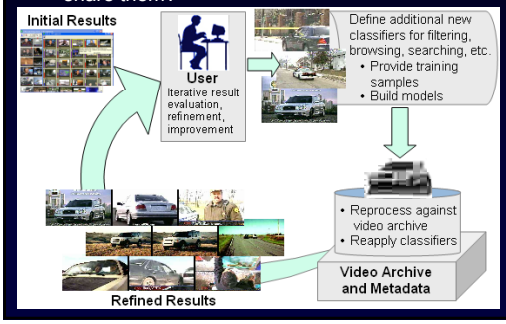


Human Computer Interaction: Better AV Search Interfaces

- Principles for dealing with errorful analysis
- Relevance Feedback, active learning, incremental learning
- Collective learning (ESP Game)
- User profiles
  - Never happens beyond a trivial way
  - No long term users in research, systems change too fast

Active Learning for Video Information Extraction

Users create their own concept classifiers for search and share them?



Carnegie Mellon

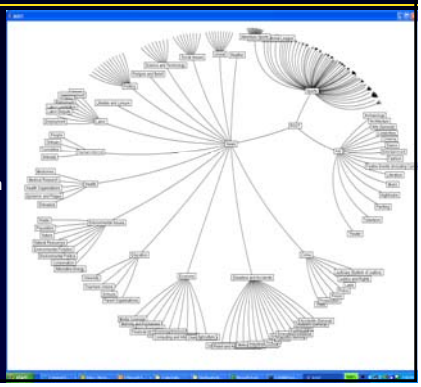
Combining World Knowledge with AV Analysis

- Better understanding for better search
  - What is the program (content)
    - What is new, what is redundant
  - What does the user want
- What summaries are useful, effective and can be automatically generated
  - Keyframes, storyboard, skims, text summaries, …
- In text analysis, ontologies (i.e. world knowledge) have NOT been a big success story!
  - This doesn't work at any reasonable scale

Carnegie Mellon

Ontologies Won't Solve These Problems

- Hand-constructed
  - As good as person building it
- Brittle with missing and extraneous information
- Don't match the data in the archive
- Can't deal with the high error rates from AV analysis



Event Search (R. Jain)

Richer analysis of importance and temporal aspects

- Calendar, blog, webpage, both "push" and "pull" technology

Not clear exactly how it should work

Carnegie Mellon

Exploiting the Next Level of Scale

- GB connectivity everywhere
- Terabyte personal storage
- Device size, mobility/location, ubiquity
- Partnering with Industry
  - to get large data sets,
  - long term studies,
  - fielded applications
  - with appropriate hardware

Carnegie Mellon

Immediate challenges

- Show generality of cross-modal approach over several domains
  - Show benefit of web-based image/video and annotations
- Can concept ***CLASSES*** work with less analysis
  - Maybe: People, objects, setting, activities
- Show benefit of using dynamic nature of video
  - Events
- Show that an ontology can help
  - How to connect logical relations to uncertain detectors?
  - Good luck!
- Show that 'iconological' concepts can be detected
  - E.g. funny, sarcastic, cozy, …

Immediate challenges

- How to leverage concept detectors for search?
  - How to present detectors to users?
  - How to select the correct detectors?
  - How to combine concept detectors?
  - How to combine concept selection methods?
- What if there is no text to search
  - Consider home video domain for example
- How to balance semantic coverage and anticipated performance of detectors for a specific query?

Economic Challenges

- Money (e.g. from advertising like Google)
- Security surveillance (industrial, public spaces)
- DVR, TV, cell-phones
  – large market saturated with high resolution, large screen devices, high bandwidth connections and nothing worthwhile to display

Automatic analysis for search concepts
– At least for "computer assisted" search

Social Challenges

- Personal recordings
  - Multimedia phones (with camera, video recorder, GPS)
    - Seagate: Terabyte on the person in 5+ years
  - Digital cameras, camcorders
  - Add GPS, time, biometrics, temperature, other sensor data
- Web2.0: YouTube, Google Video, …

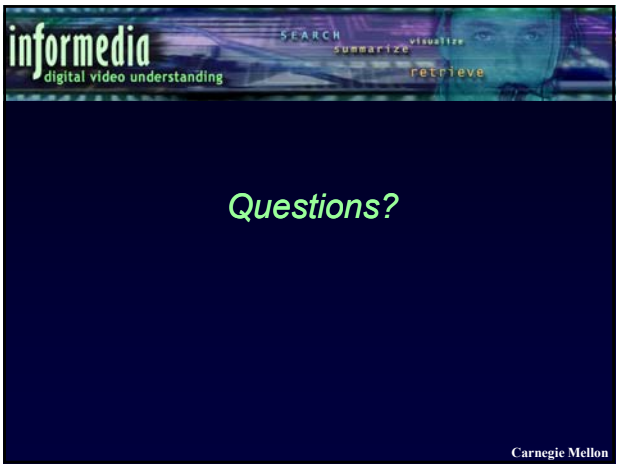
How to archive, categorize, search and leverage this flood of data

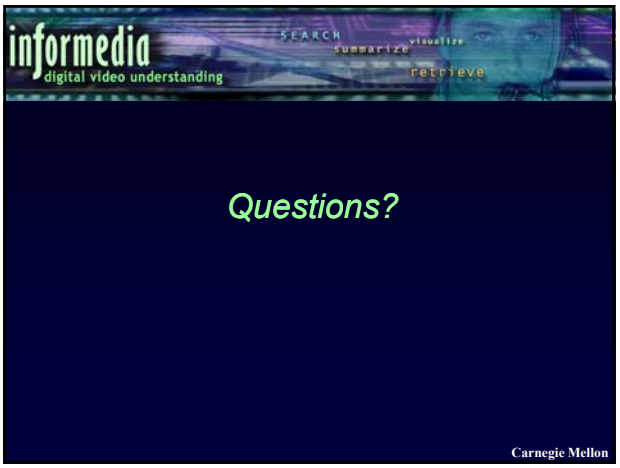
Political and Cultural Challenges

- Cultural archives
  - Preserve cultural identity in homogenizing, global world
- Access to news and information from other countries, languages
  - Broadening Perspectives
  - Communication and understanding across cultures and societies
  - Better informed citizens → more stable democracies
- Intelligence/Military Analysis
  - Foreign news imagery analysis
  - Satellite imagery, unmanned aerial vehicles
  - Web video
- Healthcare
  - Analysis of imagery
  - Combine with other sensor data
  - Longitudinal video (not just AV, but other sensors also)
  - Reduce costs
  - Unified, verifiable records

The Fundamental (BIG) Challenge

- Understanding video in general way
  - How do we see? Understand?
  - Given ***any*** video, describe it in words that are useful for search and summary
- Can computers be better than people
  - Calculators, chess, spelling (?), … AV analysis and search
- Knowing when we don't know an answer
- Mind reading
  - understand a search question in the user's mental context


informedia
digital video understanding
SEARCH
summarize
visualize
retrieve
Questions?
Carnegie Mellon


informedia
digital video understanding
SEARCH
summarize
visualize
retrieve
Questions?
Carnegie Mellon