Concept-Based Video Retrieval

Cees Snoek and Marcel Worring

with contributions by:
many

**Intelligent Systems Lab Amsterdam,
University of Amsterdam, The Netherlands**

3

The science of labeling

- To understand anything in science, things have to have a name that is recognized and is universal

naming 'categories'

naming chemical elements

naming human genome

naming rocks and minerals

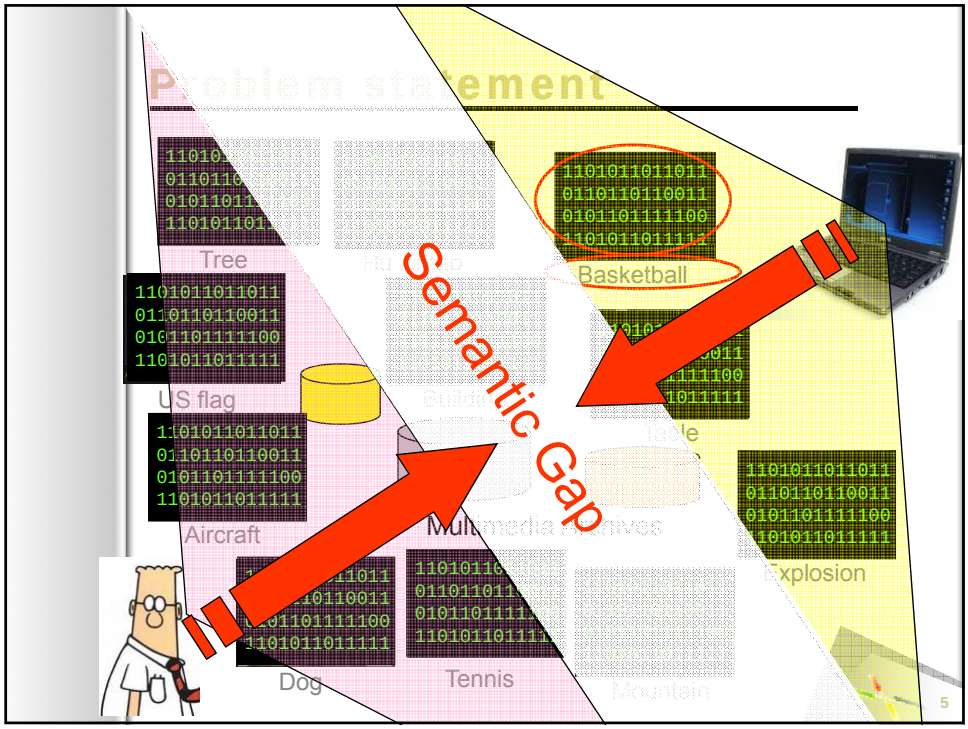
Google

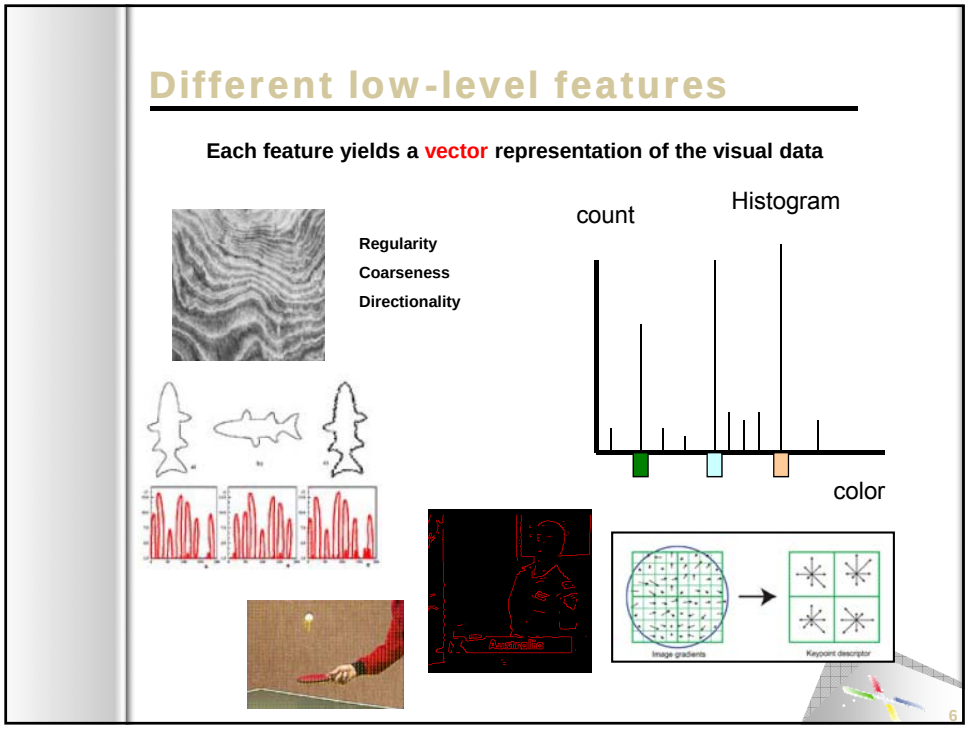
naming textual information

naming living organisms

What about naming video information?

4


Problem statement
Semantic Gap
Tree
Basketball
US flag
Aircraft
Dog
Tennis
Explosion
5

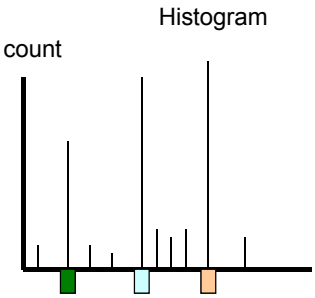

Different low-level features
Each feature yields a vector representation of the visual data
Regularity
Coarseness
Directionality
count
Histogram
color
6

Basic example: color histogram

380 pixels

640 pixels

Total 243200 pixels

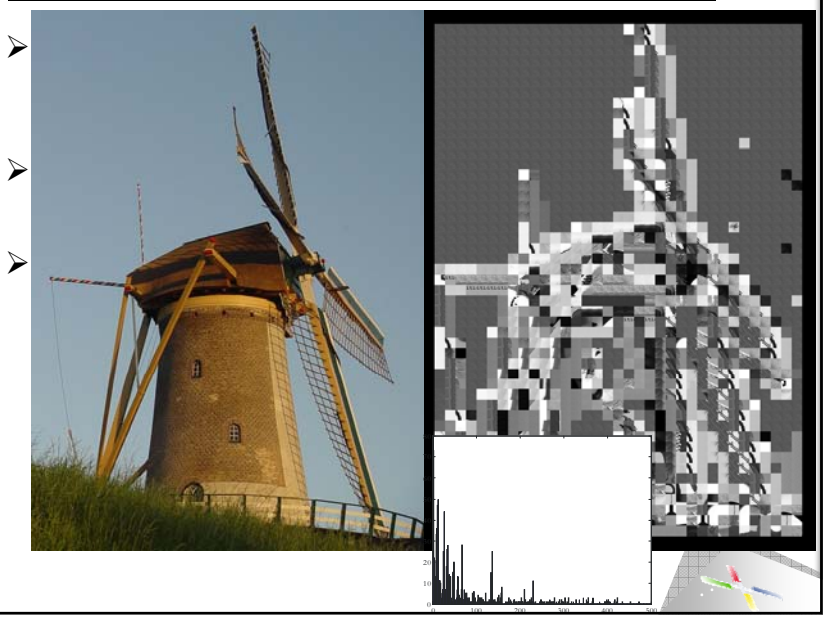


Histogram is a summary of the data summarizing in this case color characteristics

7

Leung and Malik. IJCV, 2001.
Sivic and Zisserman. ICCV, 2003.
van Gemert, PhD thesis, UvA, 2008.

Advanced example: codebook model



The goal: semantic video indexing

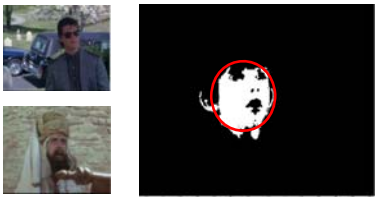
- Is the process of automatically detecting the presence of a semantic concept in a video stream

→ Airplane

9

Semantic indexing

- The computer vision approach
  - ✓ Building detectors one-at-the-time



A face detector for frontal faces

↓ 3 years later

A face detector for non-frontal faces

One (or more) PhD for every new concept

10

So how about these?


Animal


Building


Road


Beach


Boat


Graphic


People


Car


Vegetation


Overlayed Text


Studio Setting


Outdoor

And the > 1000 others ……

11

Generic concept detection in a nutshell

outdoor aircraft

Labeled examples

Feature Extraction

Supervised Learner

Training

Testing

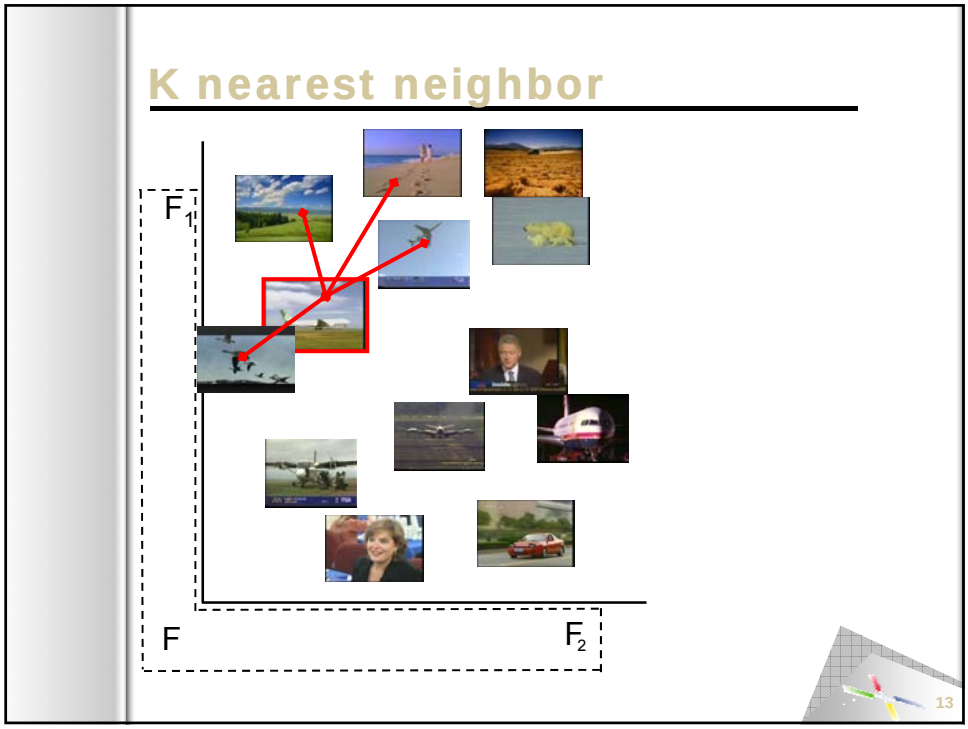

Video

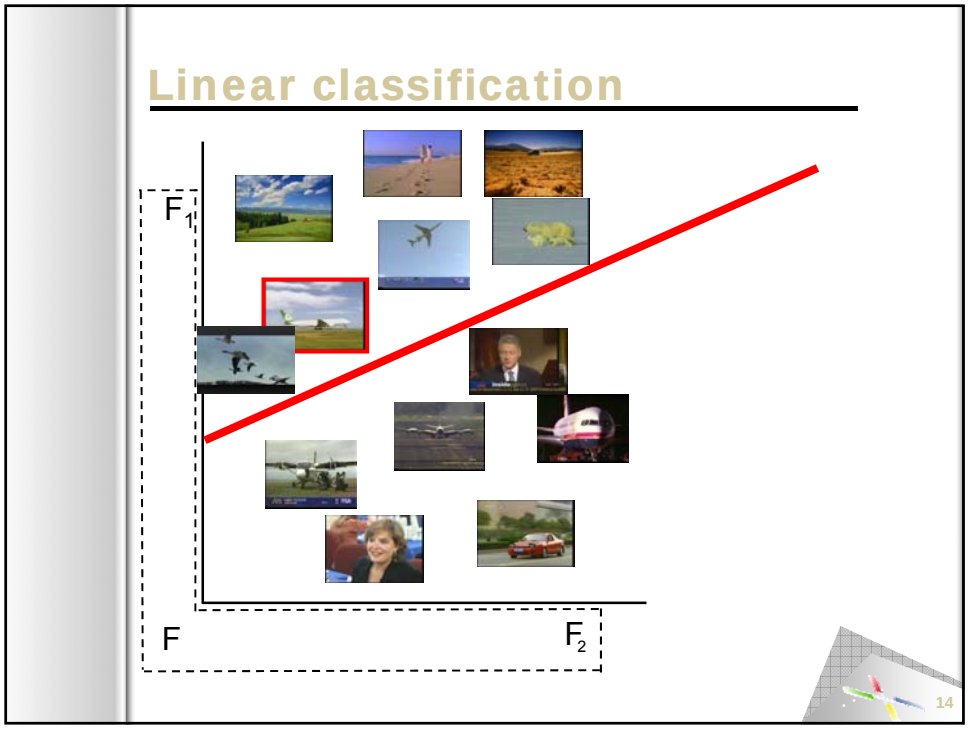
Feature Measurement

Classification

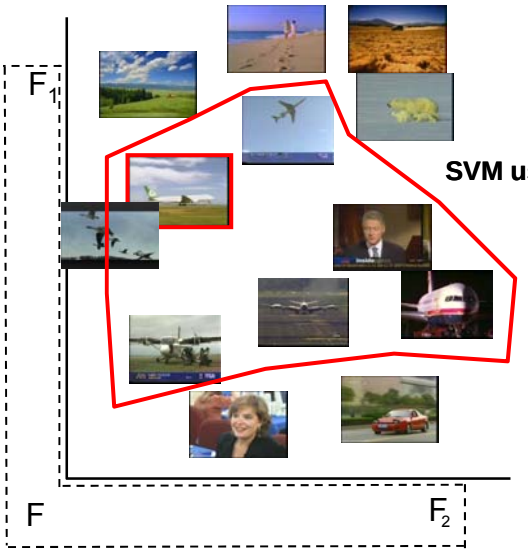
It is an outdoor aircraft probability 0.95

12


K nearest neighbor
F1
F
F2
13


Linear classification
F1
F
F2
14

Support vector machine

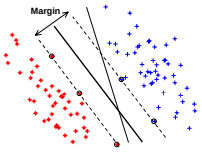


SVM usually is a good choice

15

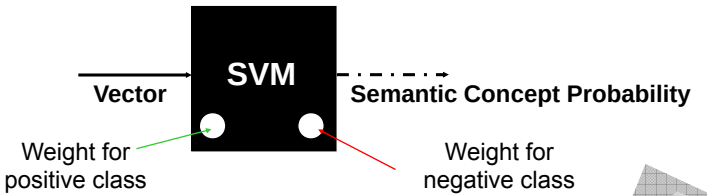
Supervised Learner

- Support Vector Machine
  - ✓ Learns from provided examples
  - ✓ Maximizes margin between two classes

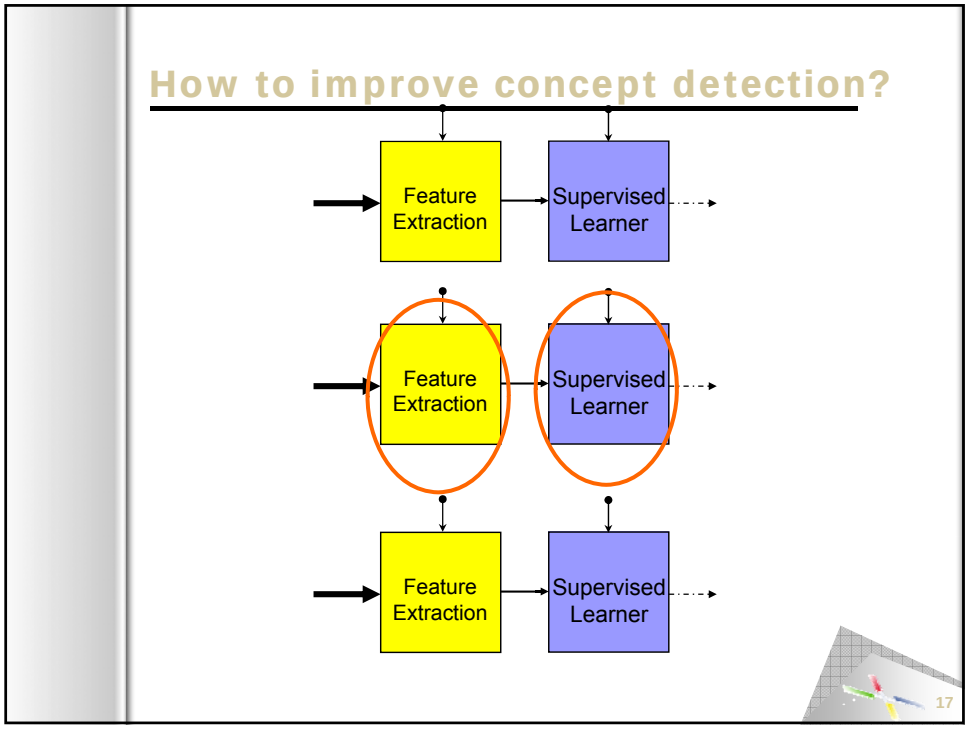


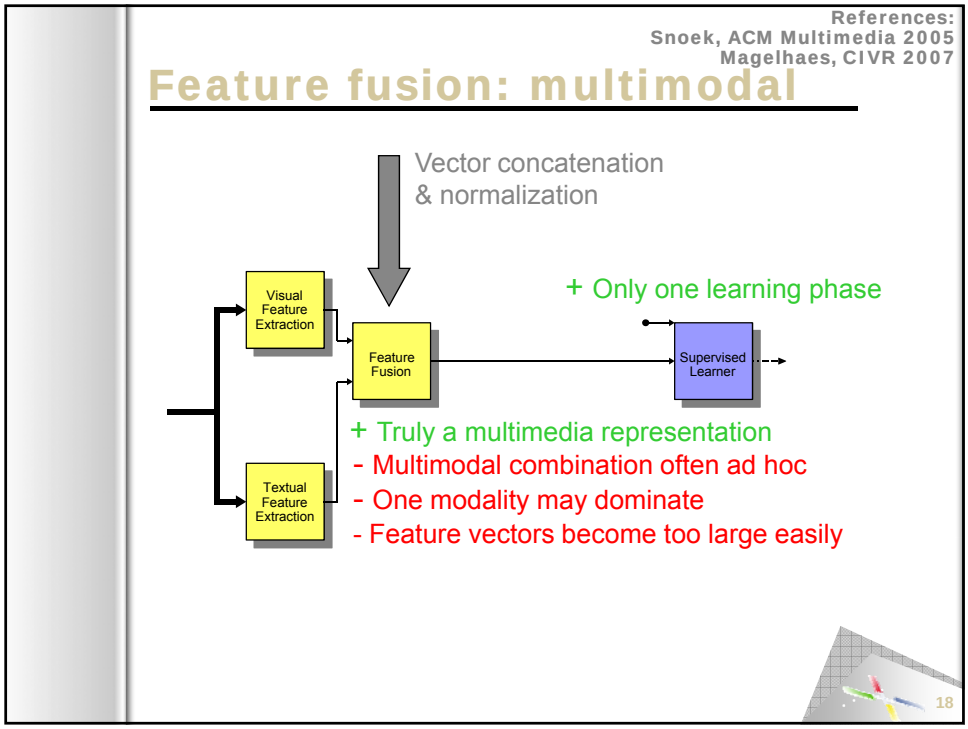
$$p(\omega|\vec{s}) = \frac{1}{1+\exp(\alpha\gamma(\vec{s})+\beta)}$$

- Depends on many parameters
  - ✓ Select best of multiple parameter combinations
  - ✓ Using cross validation



16

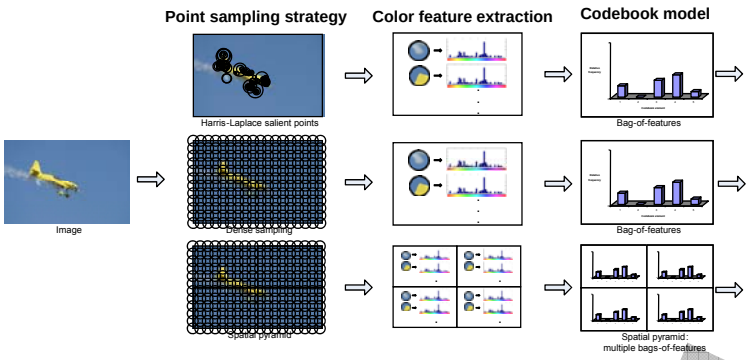

How to improve concept detection?
Feature Extraction
Supervised Learner
Feature Extraction
Supervised Learner
Feature Extraction
Supervised Learner
17


References:
Snoek, ACM Multimedia 2005
Magelhaes, CIVR 2007
Feature fusion: multimodal
Vector concatenation
& normalization
+ Only one learning phase
Visual Feature Extraction
Feature Fusion
Supervised Learner
Textual Feature Extraction
+ Truly a multimedia representation
- Multimodal combination often ad hoc
- One modality may dominate
- Feature vectors become too large easily
18

References:
van de Sande, CIVR 2008

Feature fusion: unimodal

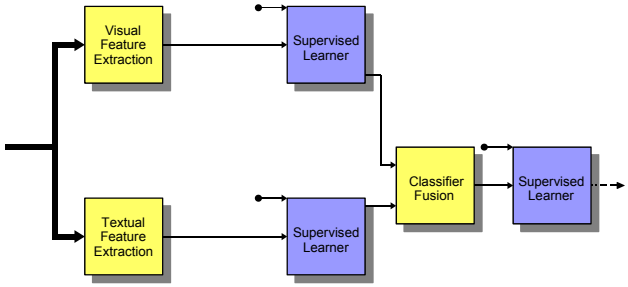
+ Codebook model reduces dimensionality
- Combination still ad hoc
- One feature may dominate



References:
Wu, ACM Multimedia 2004
Snoek, ACM Multimedia 2005

Classifier fusion: multimodal

+ Focus on modality strength
+ Fusion in semantic space

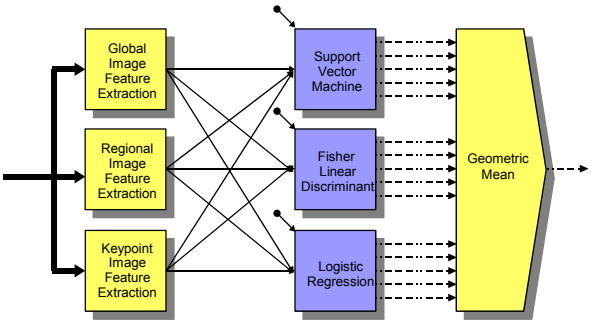


- Expensive in terms of learning effort
- Possible loss of feature space correlation

20

References:
Snoek, TRECVID 2006
Wang, ACM MIR 2007

Classifier fusion: unimodal



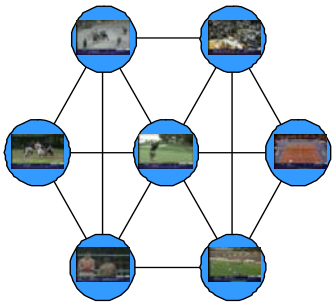
+ Aggregation functions reduce learning effort
+ Offers opportunity to use all available examples efficiently
- Linear function likely to be sub-optimal

21

References: IBM 2003
Naphade and Huang, TMM 3(1) 2001

Modeling relations

- Exploitation of conceptual co-occurrence
  - ✓ Concepts do not occur in vacuum
  - ✓ In contrast, they are related



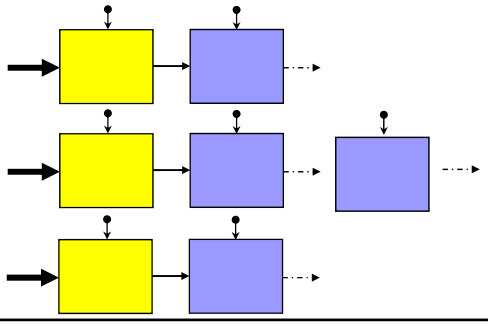
- What is sports?
  - ✓ Answer: a combination of various individual sports

22

References: IBM 2003
Qi , ACM Multimedia 2007
Liu, IEEE TMM 2008

Modeling relations

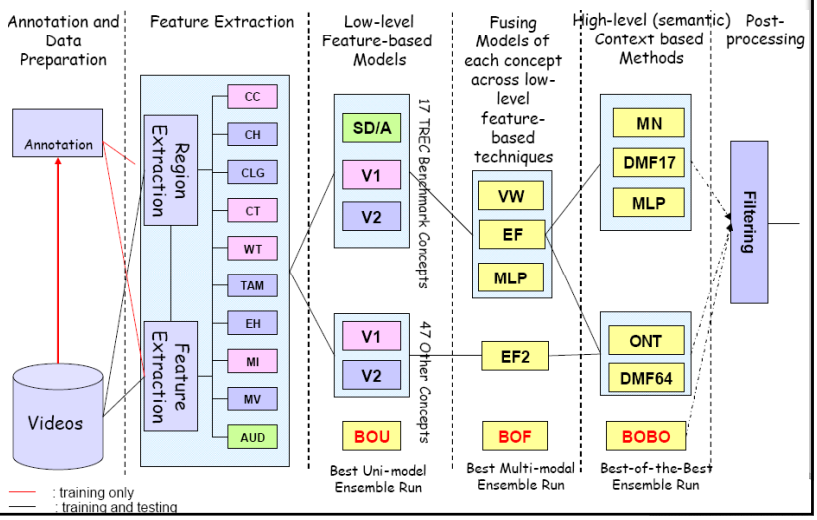
- Learning co-occurrence
  - ✓ Explicitly model relations: using graphical models
    - – Computationally complex
    - – Limited scalability
  - ✓ Implicitly learn relations: using SVM, or data mining tools
    - – Assumes classifier learns relations
    - – Suffers from error propagation

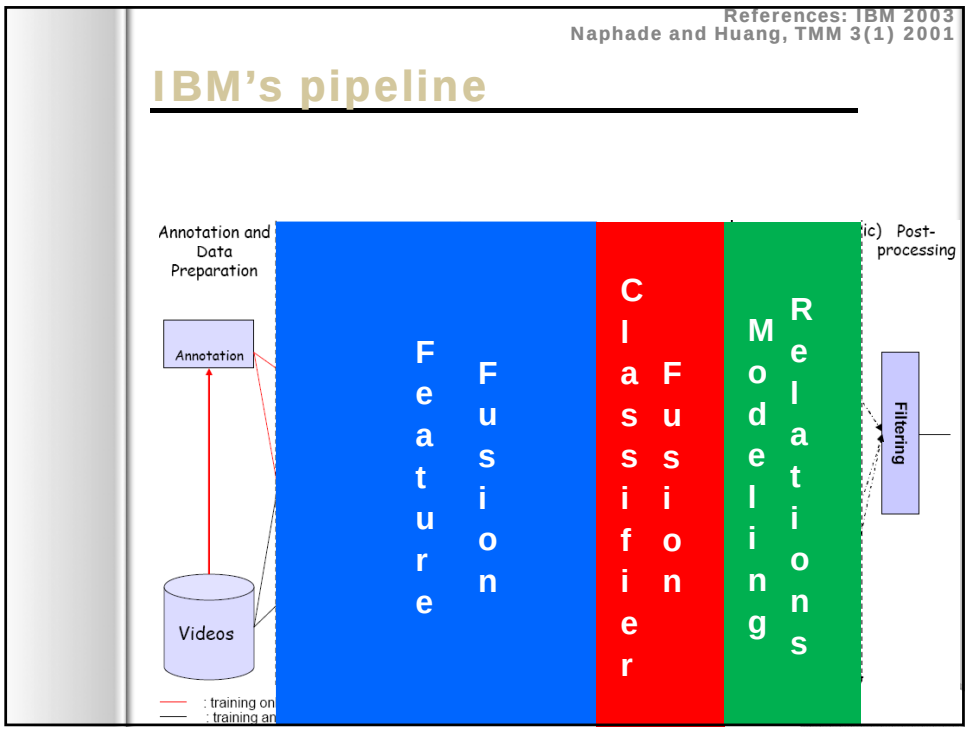


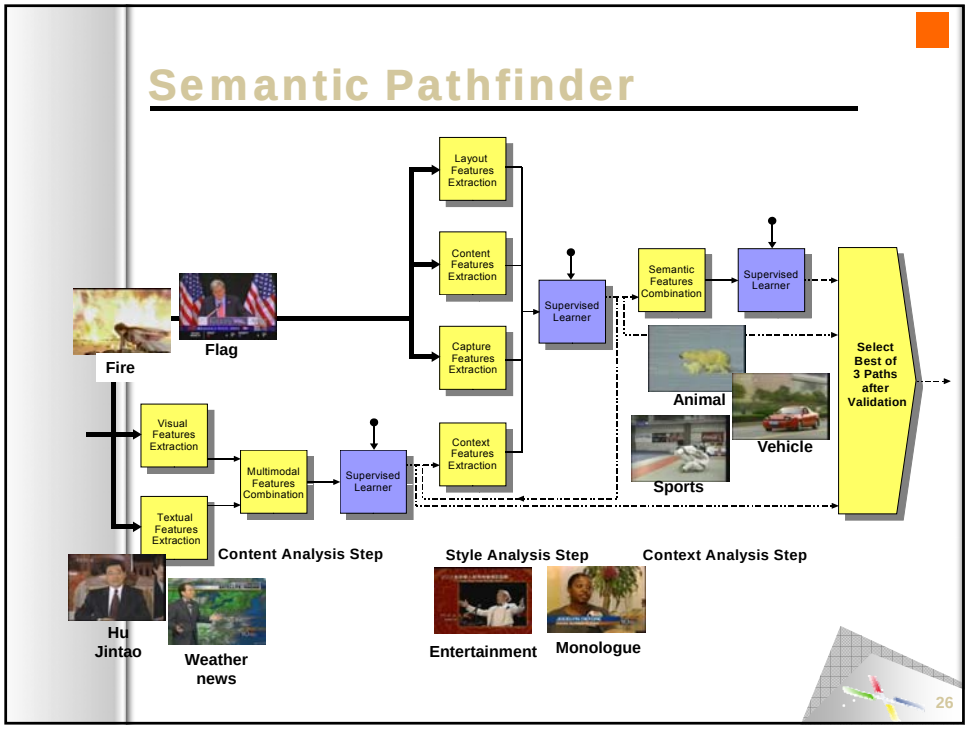
23

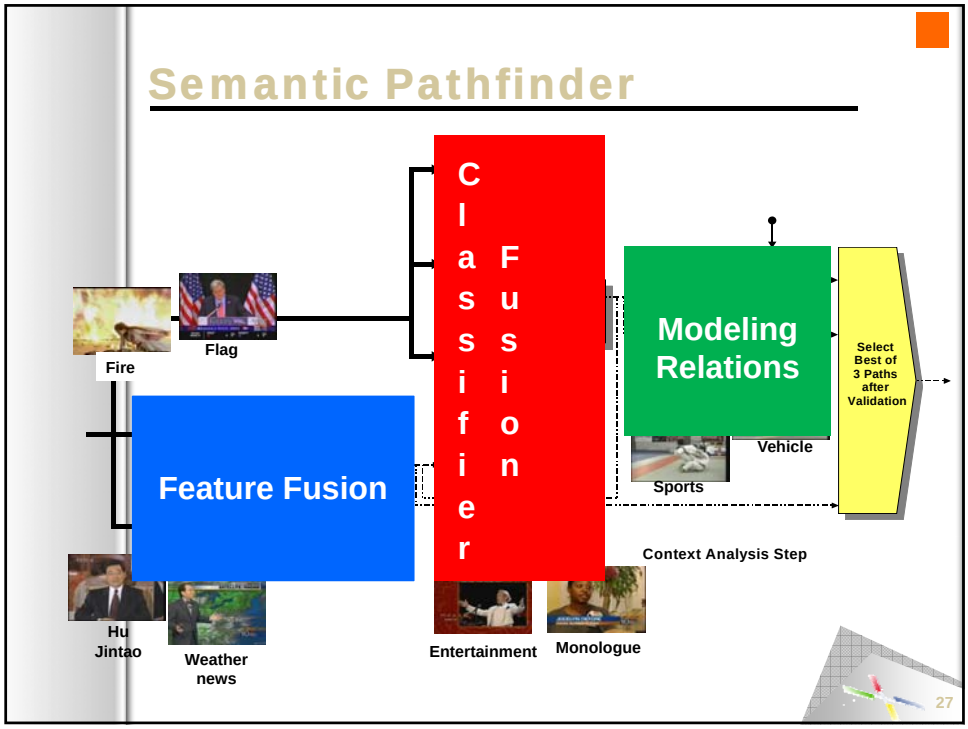
References: IBM 2003
Naphade and Huang, TMM 3(1) 2001

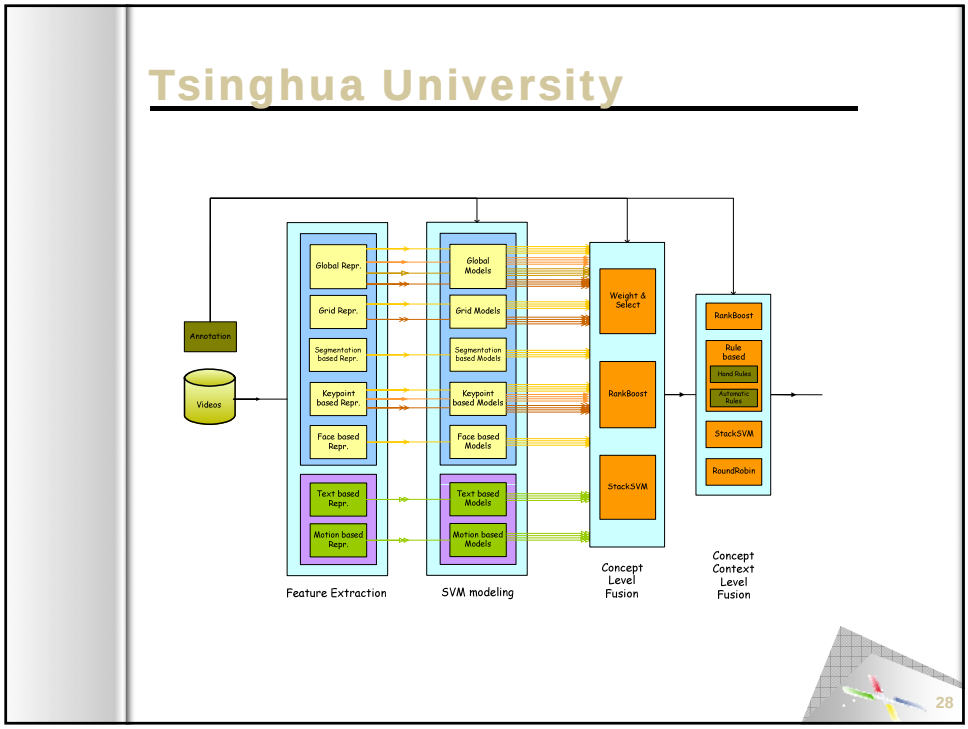
IBM's pipeline



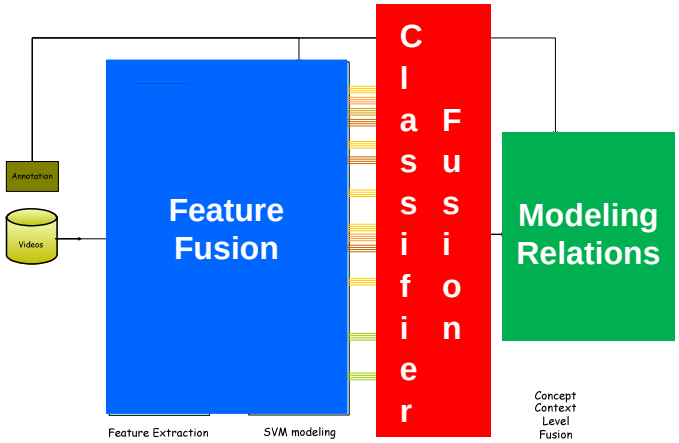

References: IBM 2003
Naphade and Huang, TMM 3(1) 2001
IBM's pipeline
Annotation and Data Preparation
Annotation
Videos
Feature Fusion
Classifier Fusion
Modeling Relations
Post-processing
Filtering


Semantic Pathfinder
Layout Features Extraction
Content Features Extraction
Capture Features Extraction
Context Features Extraction
Supervised Learner
Semantic Features Combination
Supervised Learner
Select Best of 3 Paths after Validation
Fire
Flag
Animal
Vehicle
Sports
Visual Features Extraction
Multimodal Features Combination
Supervised Learner
Textual Features Extraction
Content Analysis Step
Style Analysis Step
Context Analysis Step
Hu Jintao
Weather news
Entertainment
Monologue
26


Semantic Pathfinder
Classifier Fusion
Modeling Relations
Select Best of 3 Paths after Validation
Fire
Flag
Feature Fusion
Vehicle
Sports
Context Analysis Step
Hu Jintao
Weather news
Entertainment
Monologue
27

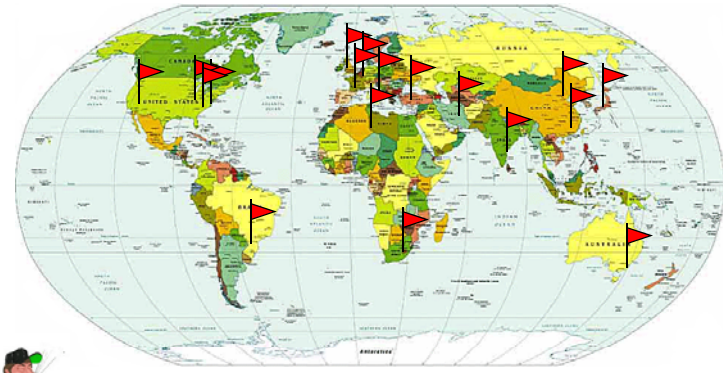

Tsinghua University
Annotation
Videos
Global Repr.
Grid Repr.
Segmentation based Repr.
Keypoint based Repr.
Face based Repr.
Text based Repr.
Motion based Repr.
Global Models
Grid Models
Segmentation based Models
Keypoint based Models
Face based Models
Text based Models
Motion based Models
Weight & Select
RankBoost
StackSVM
RankBoost
Rule based
Hand Rules
Automatic Rules
StackSVM
RoundRobin
Feature Extraction
SVM modeling
Concept Level Fusion
Concept Context Level Fusion
28

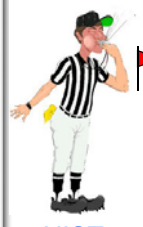
Tsinghua University



29

Fragmented research efforts…





- Video analysis researchers
  - ✓Until 2001 everybody defined her or his own concepts
  - ✓Using specific and small data sets
  - ✓Hard to compare methodologies
- Since 2001 worldwide evaluation by NIST

NIST

30

NIST TRECVID benchmark

anno 2001

- Benchmark objectives
  - ✓ Promote progress in video retrieval research
  - ✓ Provide common dataset (shots, recognized speech, key frames)
  - ✓ Use open, metrics-based evaluation
- Large international field of participants

Carnegie Mellon IBM TU Delft FX PAL MIT Microsoft

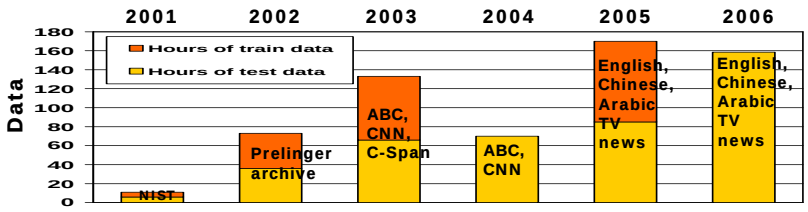
  - ✓ and the 70 others…
- Currently the de facto standard for evaluation

http://trecvid.nist.gov/

31

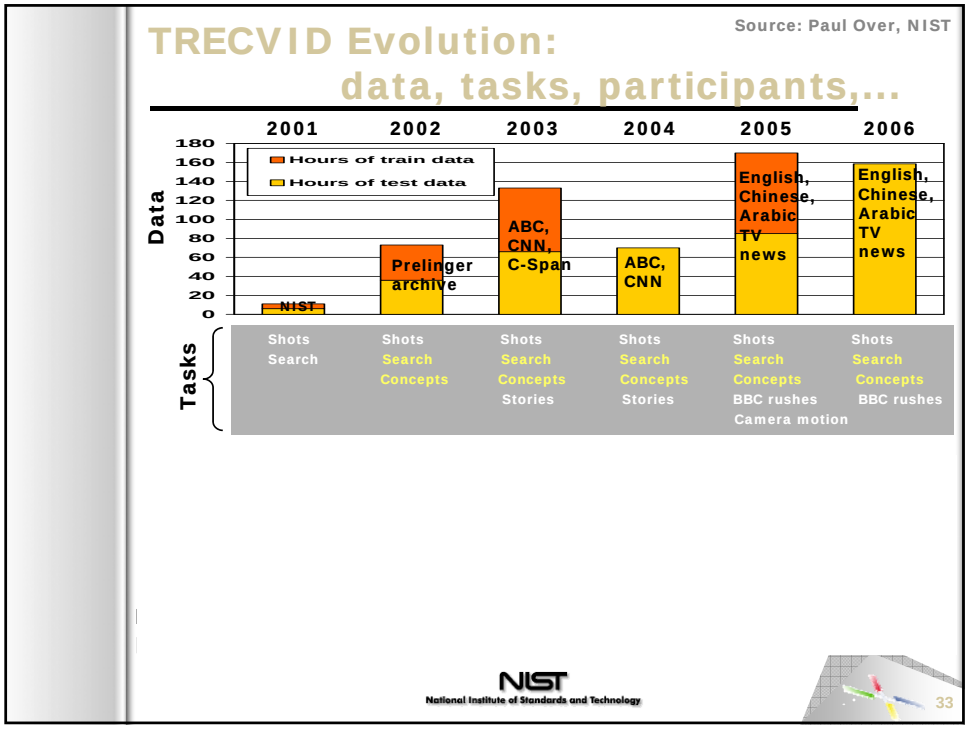
TRECVID Evolution: data, tasks, participants,...

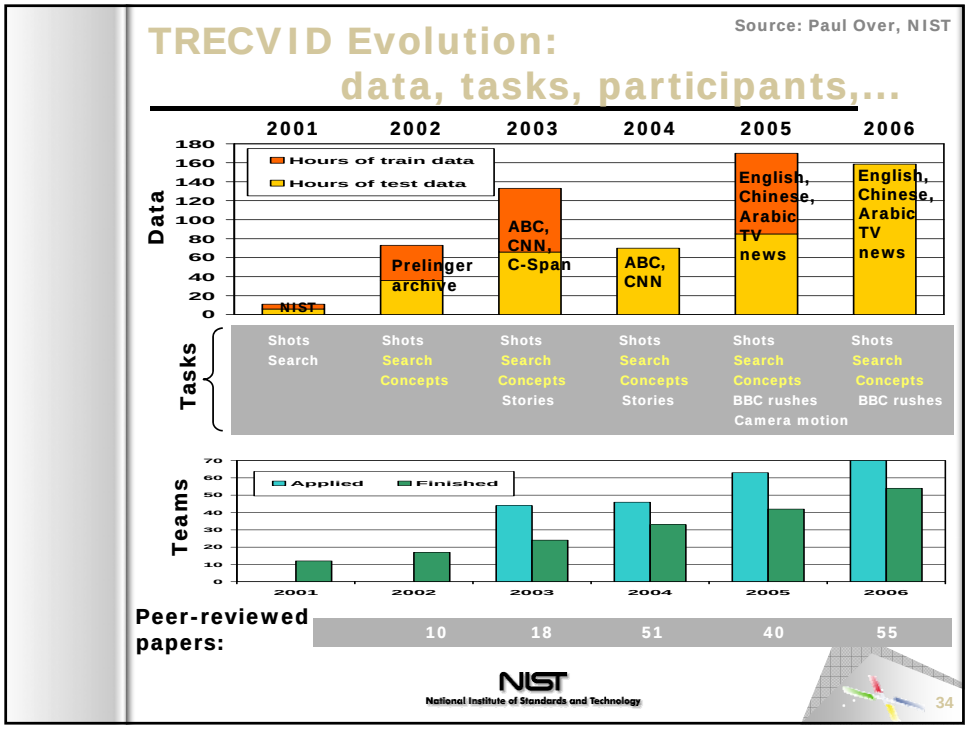
Source: Paul Over, NIST



NIST National Institute of Standards and Technology

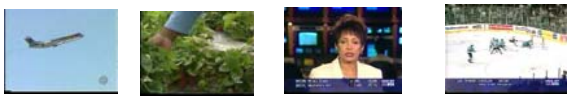
32


TRECVID Evolution:
data, tasks, participants,...
Source: Paul Over, NIST
2001
2002
2003
2004
2005
2006
Data
Hours of train data
Hours of test data
NIST
Prelinger archive
ABC, CNN, C-Span
ABC, CNN
English, Chinese, Arabic TV news
English, Chinese, Arabic TV news
Tasks
Shots
Search
Concepts
Stories
BBC rushes
Camera motion
NIST
National Institute of Standards and Technology
33


TRECVID Evolution:
data, tasks, participants,...
Source: Paul Over, NIST
2001
2002
2003
2004
2005
2006
Data
Hours of train data
Hours of test data
NIST
Prelinger archive
ABC, CNN, C-Span
ABC, CNN
English, Chinese, Arabic TV news
English, Chinese, Arabic TV news
Tasks
Shots
Search
Concepts
Stories
BBC rushes
Camera motion
Teams
Applied
Finished
Peer-reviewed papers:
10
18
51
40
55
NIST
National Institute of Standards and Technology
34

Concept detection task

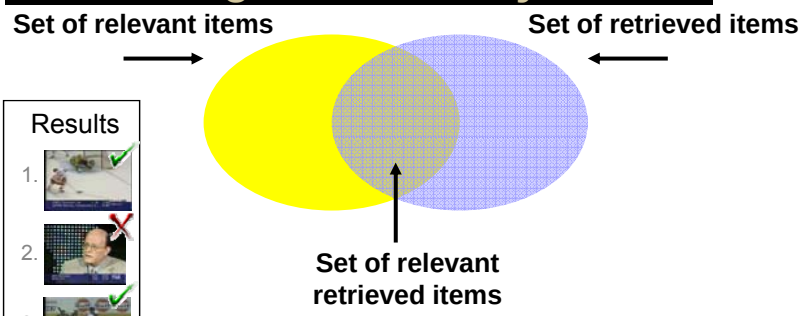
- Given:
  - ✓ a video dataset segmented into set of *S* unique shots
  - ✓ set of *N* semantic concept definitions:



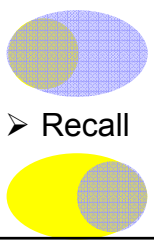
- Task:
  - ✓ How well can you detect the concepts?
  - ✓ Rank *S* based on presence of concept from *N*

35

Measuring uncertainty



- Precision
- Recall



inverse relationship

36

TRECVID evaluation measures

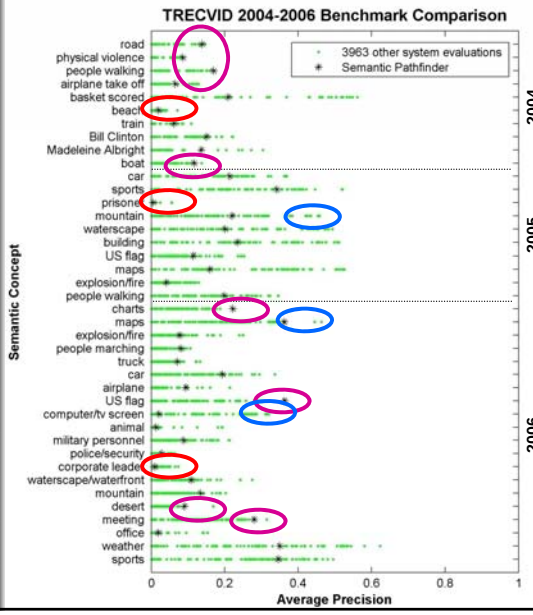
- Classification procedure
  - ✓ Training: many hours of (partly) annotated video
  - ✓ Testing: many hours of unseen video
- Evaluation measure: Average Precision
  - ✓ Combines precision and recall
  - ✓ Averages precision after every relevant shot
  - ✓ Top of the ranked list most important

$$AP = \frac{1/1 + 2/3 + 3/4 + \ldots}{\text{Total Number of correct shots}}$$

37

With the MediaMill team

Semantic Pathfinder @ TRECVID



The Good

The Bad

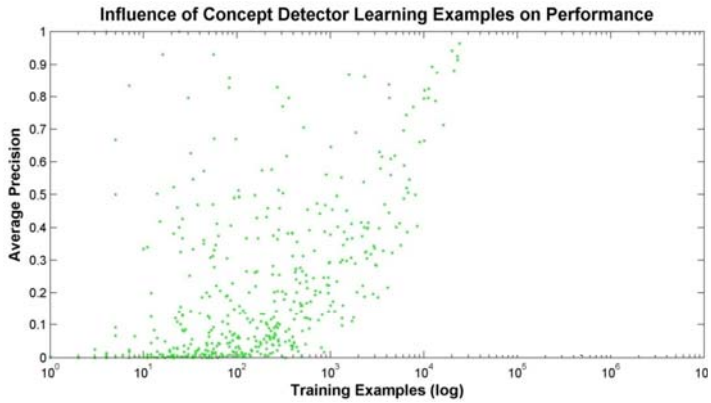
ill-defined / few examples

The Ugly

exploit TV repetition

38

491 detectors, a closer look

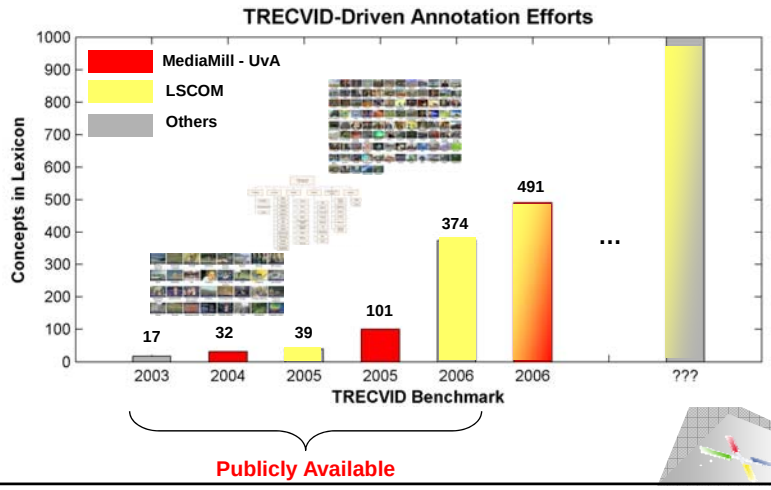


The number of labeled image examples used at training time seems decisive in concept detector accuracy.

Demo time!

Concept detector: requires examples

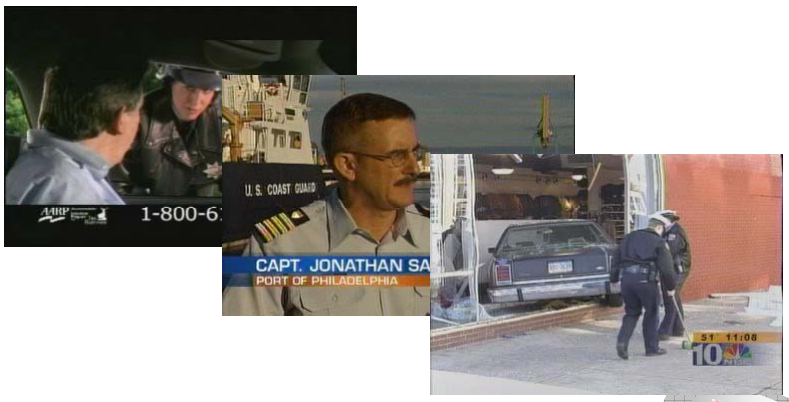
- TRECVID's collaborative research agenda has been pushing manual concept annotation efforts



41

Concept definition

- MM078-Police/Security Personnel
  - ✓ Shots depicting law enforcement or private security agency personnel.



42

References:
Christel, Informedia, 2005
Volkmer et al, ACM MM 2005

Collaborative annotation tool

TRECVID 2005

- Manual annotation by 100+ TRECVID participants
  - ✓ Incomplete, but reliable

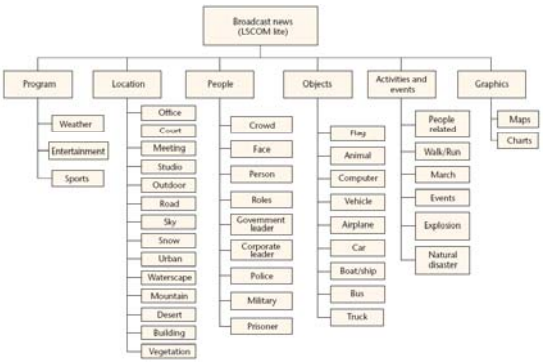
43

References:
Naphade et al, IEEE Multimedia 2006

Manual annotations: LSCOM-lite

TRECVID 2005

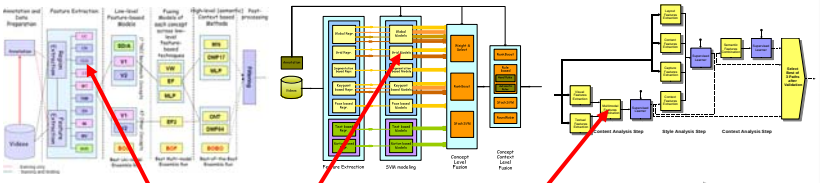
- LSCOM:
  - ✓ Large Scale Annotation for Multimedia
  - ✓ Aims for ontology of 1,000 annotated concepts
- LSCOM-Lite: annotations for 39 semantic concepts
  - ✓ Used in TRECVID 2005 and 2006



44

TRECVID Criticism

- Focus is on the final result
  - ✓ TRECVID judges relative merit of indexing methods
  - ✓ Ignores repeatability of intermediate analysis steps
- Systems are becoming more complex
  - ✓ Typically combining several features and learning methods
- Component-based optimization and comparison impossible



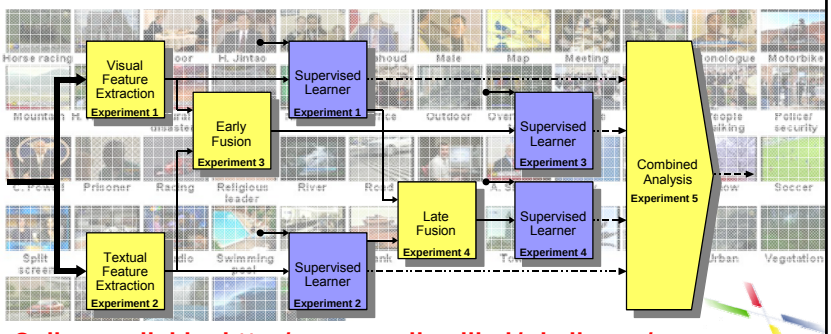
What is the contribution of these components?

45

MediaMill Challenge

- **The Challenge provides**
  - ✓ Manually annotated lexicon of 101 semantic concepts
  - ✓ Pre-computed low-level multimedia features
  - ✓ Trained classifier models
  - ✓ Five experiments
  - ✓ Baseline implementation together with baseline results
- **The Challenge allows to**
  - ✓ Gain insight in intermediate video analysis steps
  - ✓ Foster repeatability of experiments
  - ✓ Optimize video analysis systems on a component level
  - ✓ Compare and improve upon baseline

•**The Challenge lowers threshold for novice multimedia researchers**



Online available: http://www.mediamill.nl/challenge/

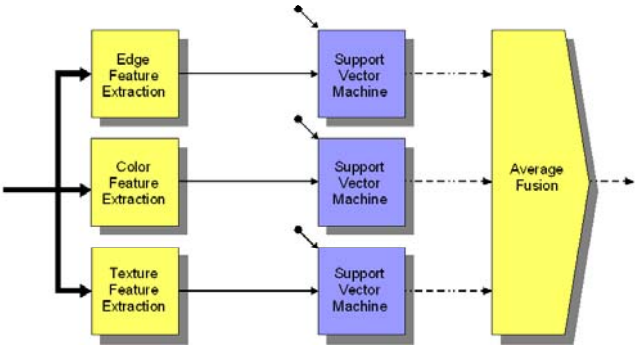
MediaMill Challenge

- Advantages
  - ✓ For research
    - – People can focus on the experiment for which they have the expertise without having to do all the processing
      - Pure computer vision
      - Pure natural language processing
      - Pure machine learning
      - ……
  - ✓ For education
    - – Students can do
      - large scale experiments
      - compare themselves to each other
      - …… and to the state-of-the-art

47

Columbia374

- Baseline for 374 concept detectors
  - ✓ Focus is on visual analysis experiments



Online available:
http://www.ee.columbia.edu/ln/dvmm/columbia374/

48

Case study

FABCHANNEL LIVE CONCERT VIDEOS

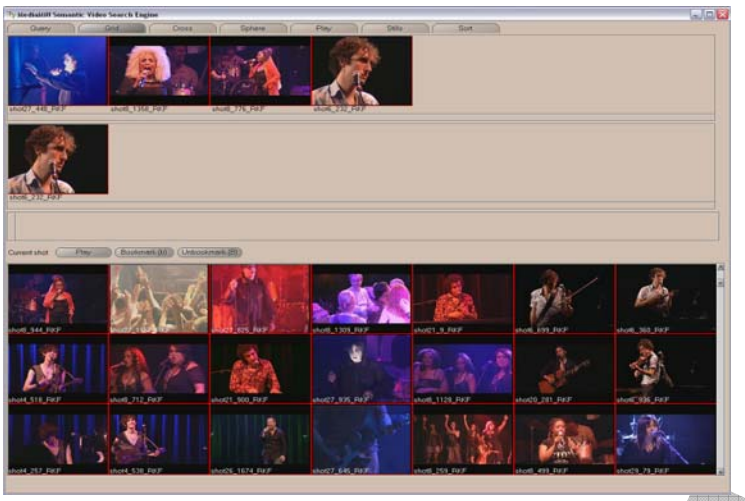
Fabchannel.com

- Fabchannel narrowcasts concerts from Amsterdam Paradiso and Melkweg venues
  - ✓ Currently +/- 700 concerts online

- Fabchannel request
  - ✓ What can you do with 45 hours of live concerts?
- Answer:
  - ✓ Let's try the semantic pathfinder to detect concert concepts

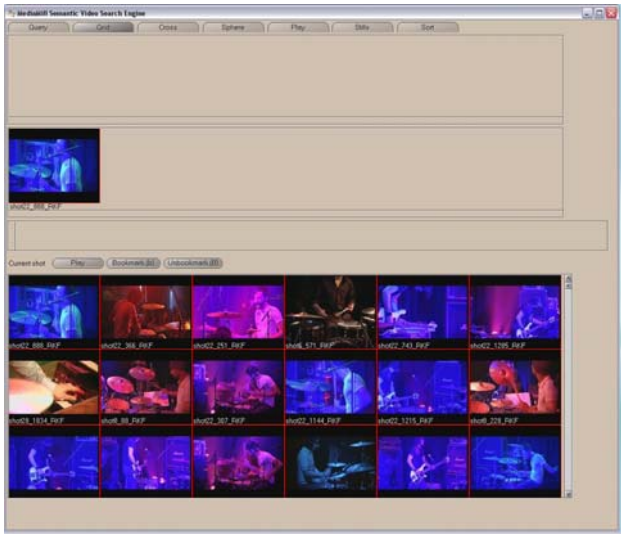
FABCHANNEL LIVE CONCERT VIDEOS

Results for singer



50

Results for drummer



51

Conclusions

- An international community is building a bridge to narrow the semantic gap
  - ✓ Currently detects more than 500 concepts in broadcast video
  - ✓ Generalizes outside news domain
- Important lessons
  - ✓ No superior method for all concepts exists,
  - ✓ Best to learn optimal approach per concept
  - ✓ Best methods cover variation in features, classifiers, and concepts

52

Concept detection challenges

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

53

Using concept detectors

- "We are now seeing researchers starting to use the confidence values from concept detectors, within the shot retrieval process and this appears to be the roadmap for future work in this area."
  - ✓ Alan Smeaton, Information Systems, 32(4):545-559, 2007

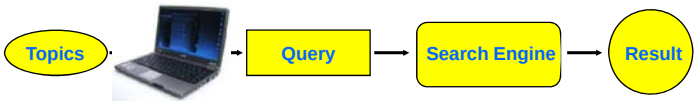
54

Measure concept detector influence

- Hypothesis 1:
  - ✓ *Increasing the number of concept detectors in a lexicon improves video retrieval accuracy.*
- Hypothesis 2:
  - ✓ *Combining concept detectors from a lexicon improves video retrieval accuracy.*

55

TRECVID automatic search task



- Automatically solve search topic
- Return 1,000 ranked shot-based results
- Evaluate using Average Precision

- TRECVID 2005
  - ✓ 85 hrs test set – Chinese, Arabic, English TV News
  - ✓ 24 search topics

Topic examples

Find shots of one or more helicopters in flight.

Find shots of a hockey rink with at least one of the nets fully visible from some point of view.

Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people

Find shots of a group including at least four people dressed in suits, seated, and with at least one flag.

57

57

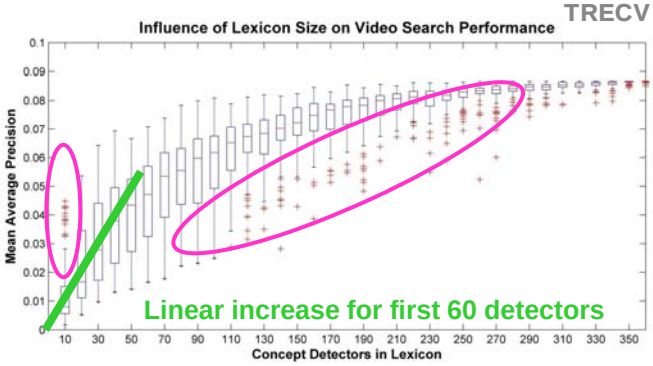
Influence of lexicon size

- Lexicon = 363 machine learned concept detectors
- Procedure
  1. Set bag size *B* = 10;
  2. Select random bag of *B* detectors from lexicon
  3. Determine maximum performance for each search topic
  4. ***B+=10;***
  5. Go back to step 2.
- Repeat the process 100 times
  - ✓ Reduces random positive and negative effects

58

Influence of lexicon size

TRECVID 2005

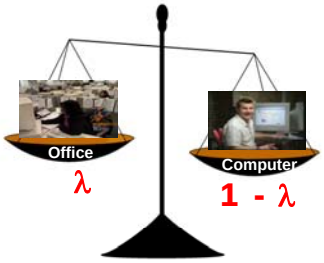


- Size matters
  - ✓ Lexicon of 150 detectors comes close to maximum performance
- Some detectors perform well for specific topics
  - ✓ Tennis game detector for "find two visible tennis players"
- Substantial number of detectors not accurate enough yet
  - ✓ Only small increase when more than 70 detectors are used

59

Influence of detector combination

- How to combine multiple detectors?
  - ✓ Experiment: pair-wise oracle fusion



- Improvement for 20 out of 24 topics
- Increase per topic as high as 89%
- Overall increase 10%

60

Typical results

Find shots of a graphic map of Iraq, location of Bagdhad marked - not a weather map.

Best:  Maps

+ 2nd Best:  Overlayed Text

How to select relevant detectors automatically?

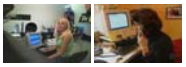
Find shots of George Bush entering or leaving a vehicle (e.g., car, van, airplane, helicopter, etc) (he and vehicle both visible at the same time)

 rocket propelled grenades

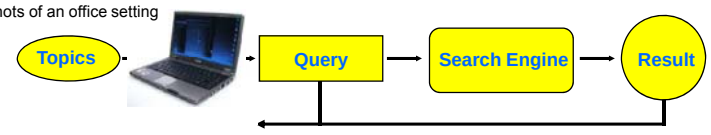
+  Iyad Allawi ?

61

Problem statement



Find shots of an office setting

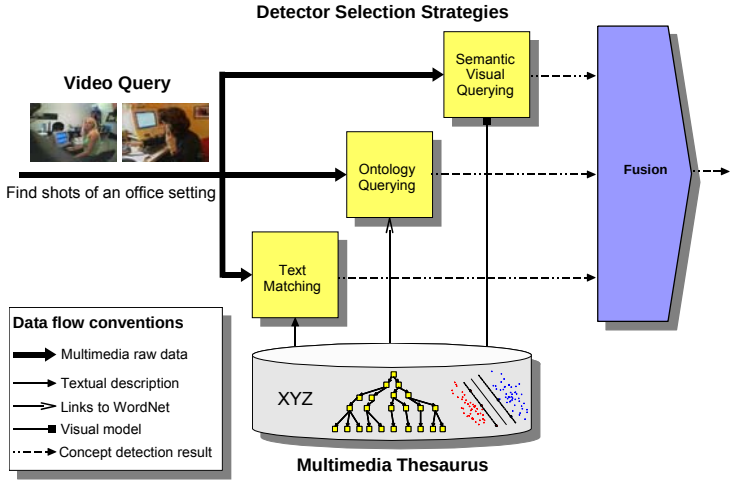


- How to translate query topic to concept detectors?

62

With B. Huurnink / M. de Rijke
L. Hollink / G. Schreiber
M. Worring

Detector selection strategies

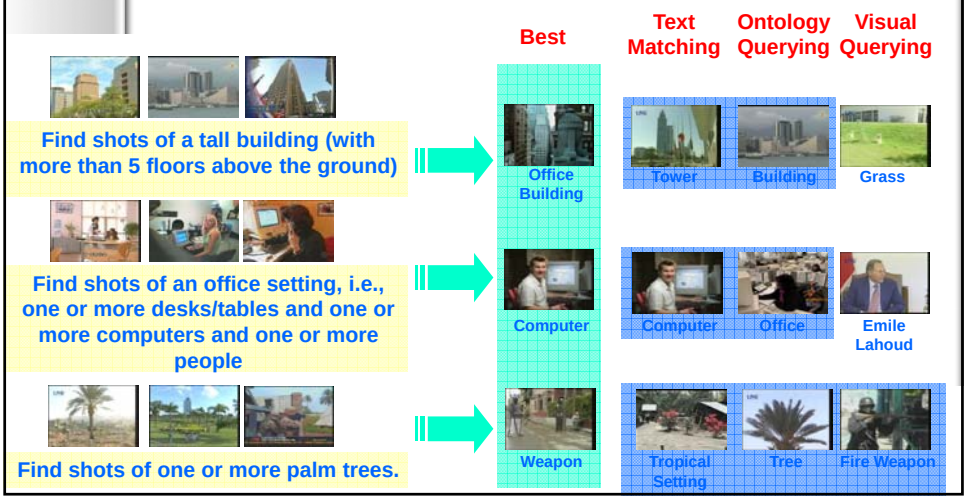


63

Influence of detector selection combi

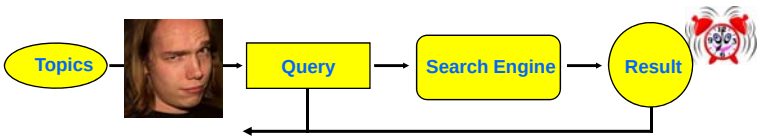
TRECVID 2005

- Individual selection strategies seem comparable
  - ✓ But, oracle combination of selection strategies pays off!



TRECVID interactive search task

- So many choices for retrieval…
  - ✓ Why not let user decide interactively?



http://trecvid.nist.gov/

65

References:
Carnegie Mellon University

'Classic' Informedia system

- First multimodal video search engine

66

References: FxPal

MediaMagic

- Focus on the story level

67

References: IBM

IBM MARVel

- A web based system

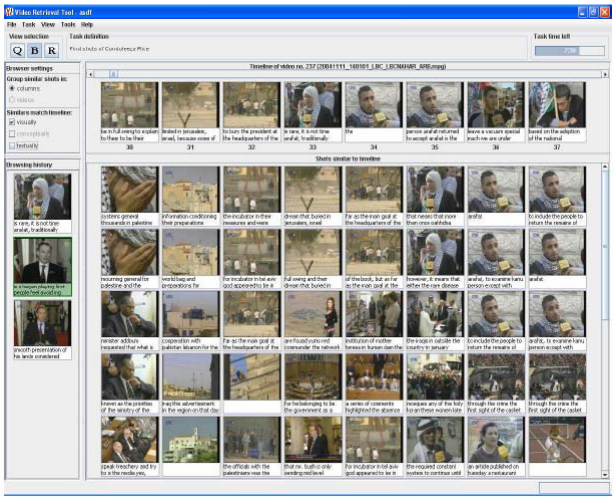
http://mp7.watson.ibm.com/marvel/

68

References:
Oulu University

Cluster-temporal browsing

- Using that result are typically similar/close in time

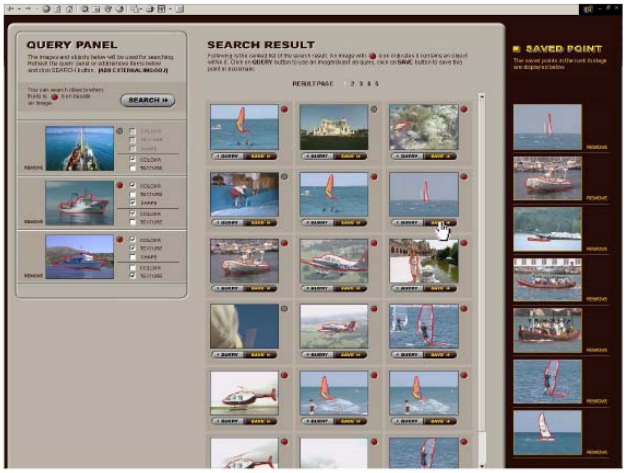


69

References:
Dublin City University

Físchlár

- Optimized for use by “real” users



70

References:
NUS & ICT-CAS

VisionGo

- Extremely fast and efficient

71

References:
Carnegie Mellon University

Extreme video retrieval

- Observation
  - ✓ Correct results are retrieved, but not optimally ranked
  - ✓ If user has time to scan results exhaustively, retrieval is a matter of watching, selecting, and sorting quickly
- Push the user to the max = very demanding!
  - ✓ ~~Rapid serial~~ visual presentation
  - ✓ Adjust browser to depth of results

72

Futuristic video retrieval

Jonathan Wang, Carnegie Mellon University

73

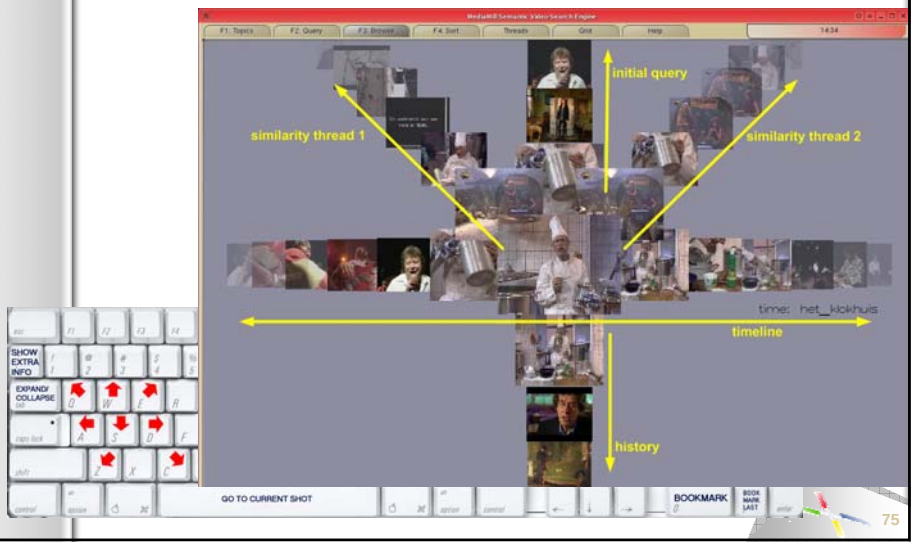
With the MediaMill team

CrossBrowsing through results

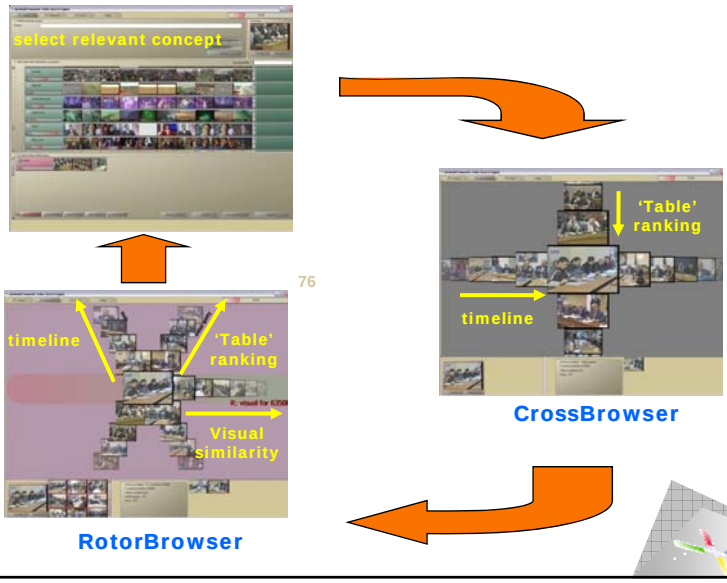
74

References:
de Rooij, CIVR 2008

ForkBrowser



Demo time!



NIST TRECVID benchmark

anno 2001

- Benchmark objectives
  - ✓ Promote progress in video retrieval research
  - ✓ Provide common dataset (shots, recognized speech, key frames)
  - ✓ Use open, metrics-based evaluation
- Large international field of participants

  - ✓ and the 70 others…
- Currently the de facto standard for evaluation

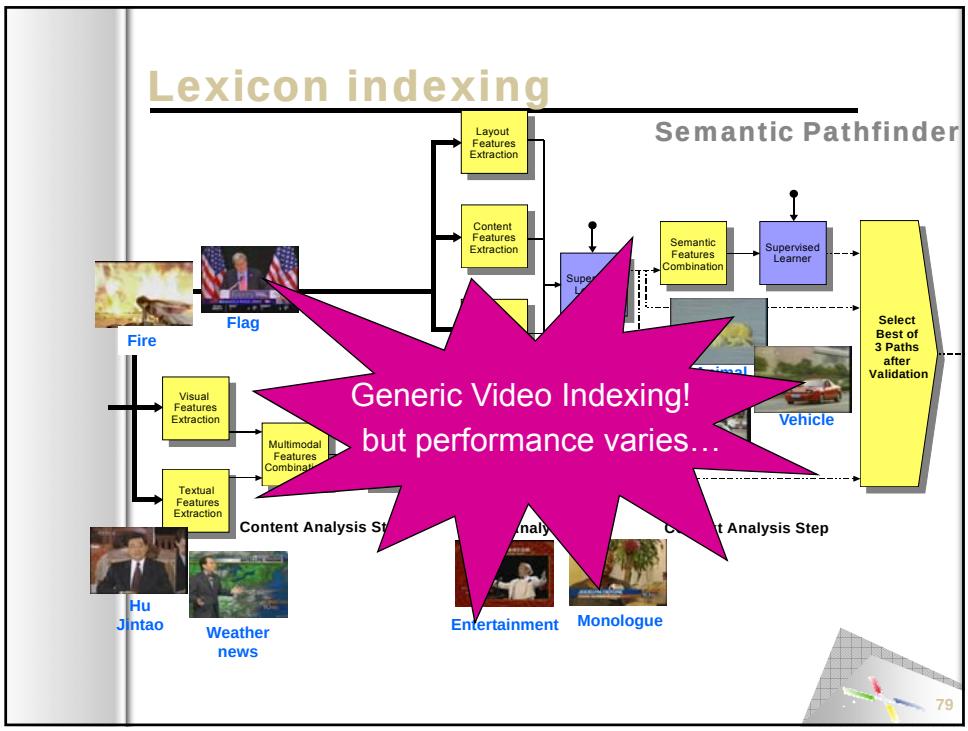
http://trecvid.nist.gov/

77

Experimental Setup

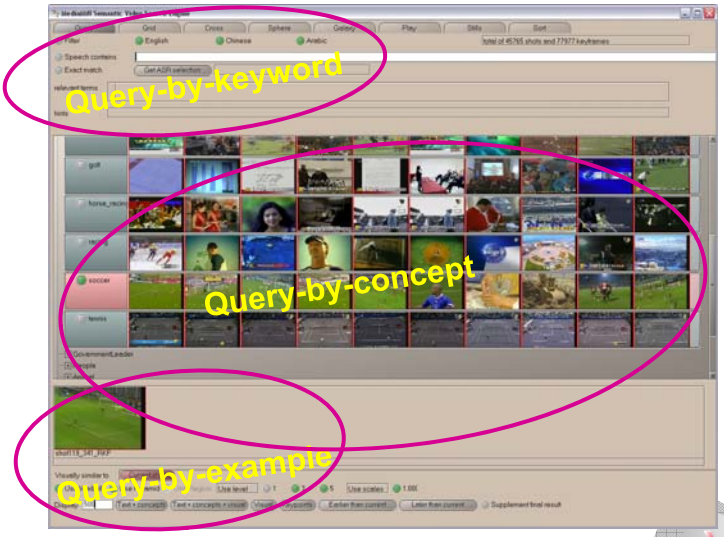
- Experiment 1
  - ✓ TRECVID 2004 (64 hrs test set – English TV News)
  - ✓ Lexicon with 32 learned concepts (where others use max. 10)
  - ✓ All other components 'standard'
- Experiment 2
  - ✓ TRECVID 2005 (85 hrs test set – Chinese,Arabic,English TV News)
  - ✓ Lexicon with 101 learned concepts (where others use max. 39)
  - ✓ Added advanced display (CrossBrowser)

78


Lexicon indexing
Semantic Pathfinder
Layout Features Extraction
Content Features Extraction
Semantic Features Combination
Supervised Learner
Select Best of 3 Paths after Validation
Visual Features Extraction
Multimodal Features Combinat
Textual Features Extraction
Content Analysis St
t Analysis Step
Generic Video Indexing!
but performance varies…
Fire
Flag
Vehicle
Hu Jintao
Weather news
Entertainment
Monologue
79


TRECVID 2004
Learned lexicon of 32 concepts
Animal
Footbal
Road
Beach
Stock Quotes
Golf
Financial Anchor
Cartoon
Building
Airplane Take Off
Boat
Graphic
People
Car
Vegetation
Overlayed Text
Basket Scored
Bill Clinton
Sporting Event
Studio Setting
Physical Violence
Train
Baseball
News Subject Monologue
Anchor
Outdoor
Ice Hockey
People Walking
Madeleine Albright
Soccer
Bicycle
Weather News
80

Query selection

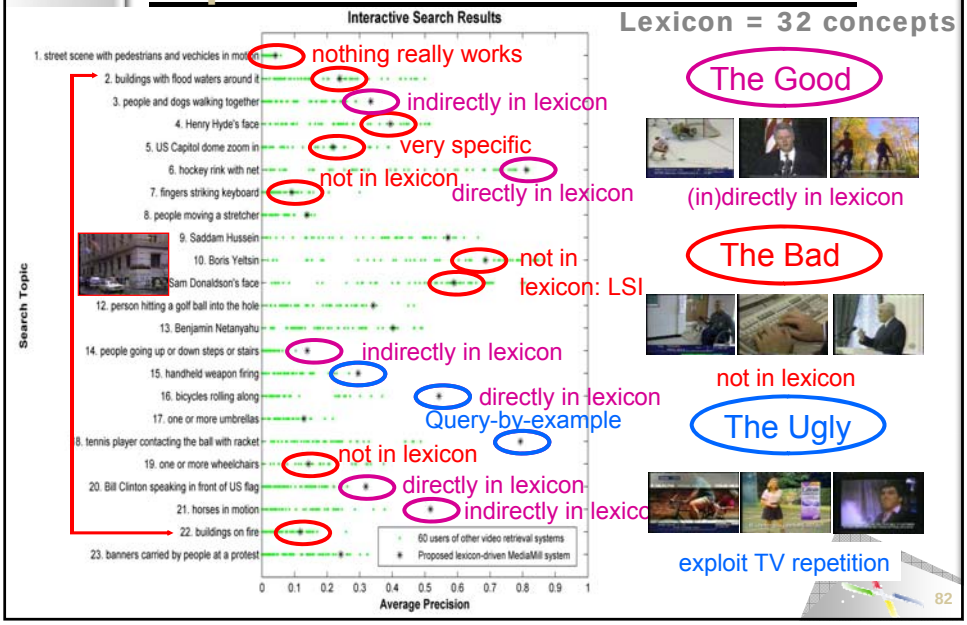


… yields a ranking of the data

81

TRECVID 2004

Experiment 1



82

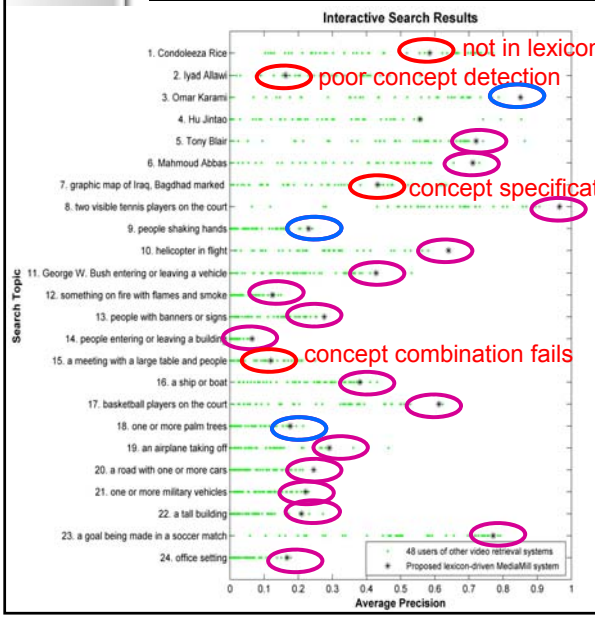
TRECVID 2005

Learned lexicon of 101 concepts

83

TRECVID 2005

Experiment 2



not in lexicon

poor concept detection

concept specification fails

concept combination fails

Lexicon = 101 concepts

The Good

Almost all topics solvable by using concept lexicon only!

The Bad

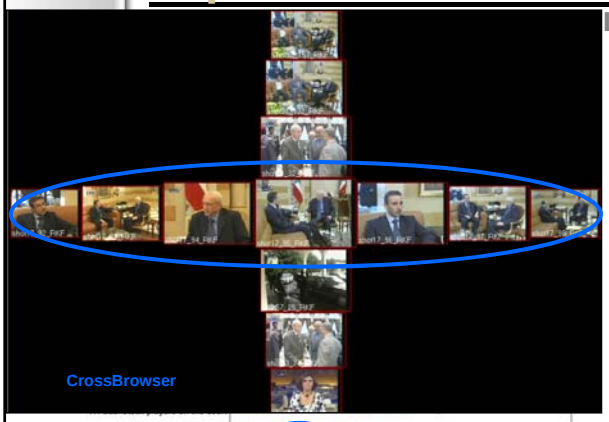
The Beautiful

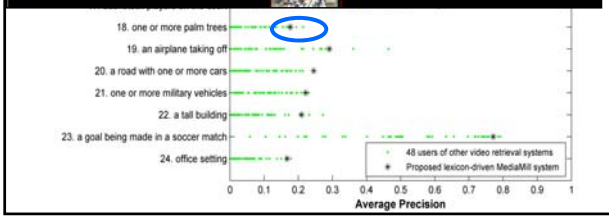
Exploit common sense!

84

TRECVID 2005

Experiment 2





Lexicon = 101 concepts

The Good

Almost all topics solvable by using concept lexicon only!

The Bad

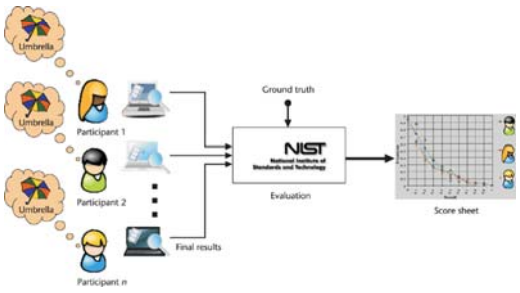
The Beautiful

Exploit common sense!

85

TRECVID criticism

- Need to realize that benchmark performance cannot be the sole criterion

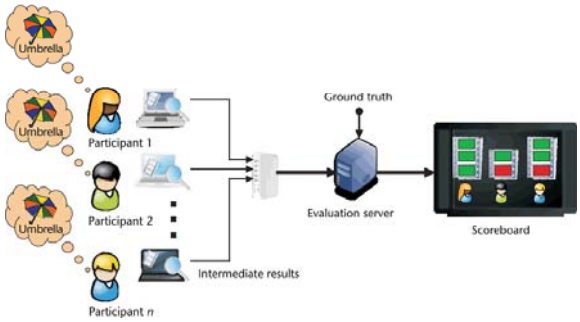


- ✓ Quality of concept detectors counts
- ✓ Experience of searcher counts
- ✓ Visualization of the interface counts
- ✓ Ease of use counts
- ✓

86

VideOlympics

- Promote multiple facets of interactive video search
  - ✓ Real-time interactive video search 'competition'
  - ✓ Simultaneous exposure of multiple video retrieval systems
  - ✓ Audience obtains complete overview of possibilities and limitations of state-of-the-art



87

Participants

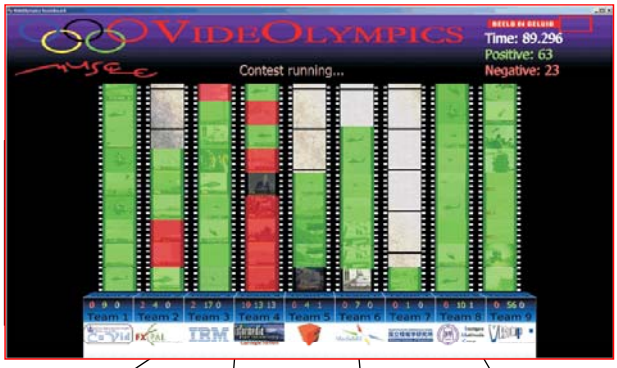
Showcase Event

88

Setup

VIDEOLYMPICS

Showcase Event



One display

TRECVID like queries
A result is submitted as soon as it is found




.....

Video trailer

VIDEOLYMPICS

http://www.VideOlympics.org

90

Conclusions

- We should use
  - ✓ Generic semantic indexing
  - ✓ Large lexicons
- For browsing
  - ✓ The large lexicons are the best entrance
  - ✓ Ranking and time are most important dimensions
- TRECVID = priceless
  - ✓ Fosters a common research agenda and international collaboration
  - ✓ Many valuable resources available online
  - ✓ Offers you an opportunity to do it yourself easily

Concept retrieval challenges

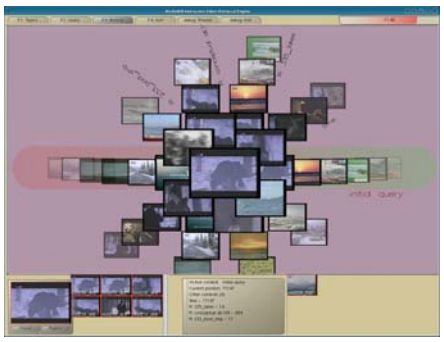
- How to leverage concept detectors for search?
  - ✓ How to select the best detectors automatically?
  - ✓ How to combine concept selection methods?
  - ✓ How to balance semantic coverage and anticipated performance of detectors for a specific query?
- How to help the user in browsing
  - ✓ How to present detectors and their uncertainty to users?
  - ✓ How to select the best detector on the fly?
  - ✓ How to present thread space in an intuitive manner?
- Generalization to unstructured domains
  - ✓ Consider YouTube domain for example
  - ✓ How to use contextual metadata?

92

Further information

MediaMill

www.MediaMill.nl



93