Concept-based Video Indexing

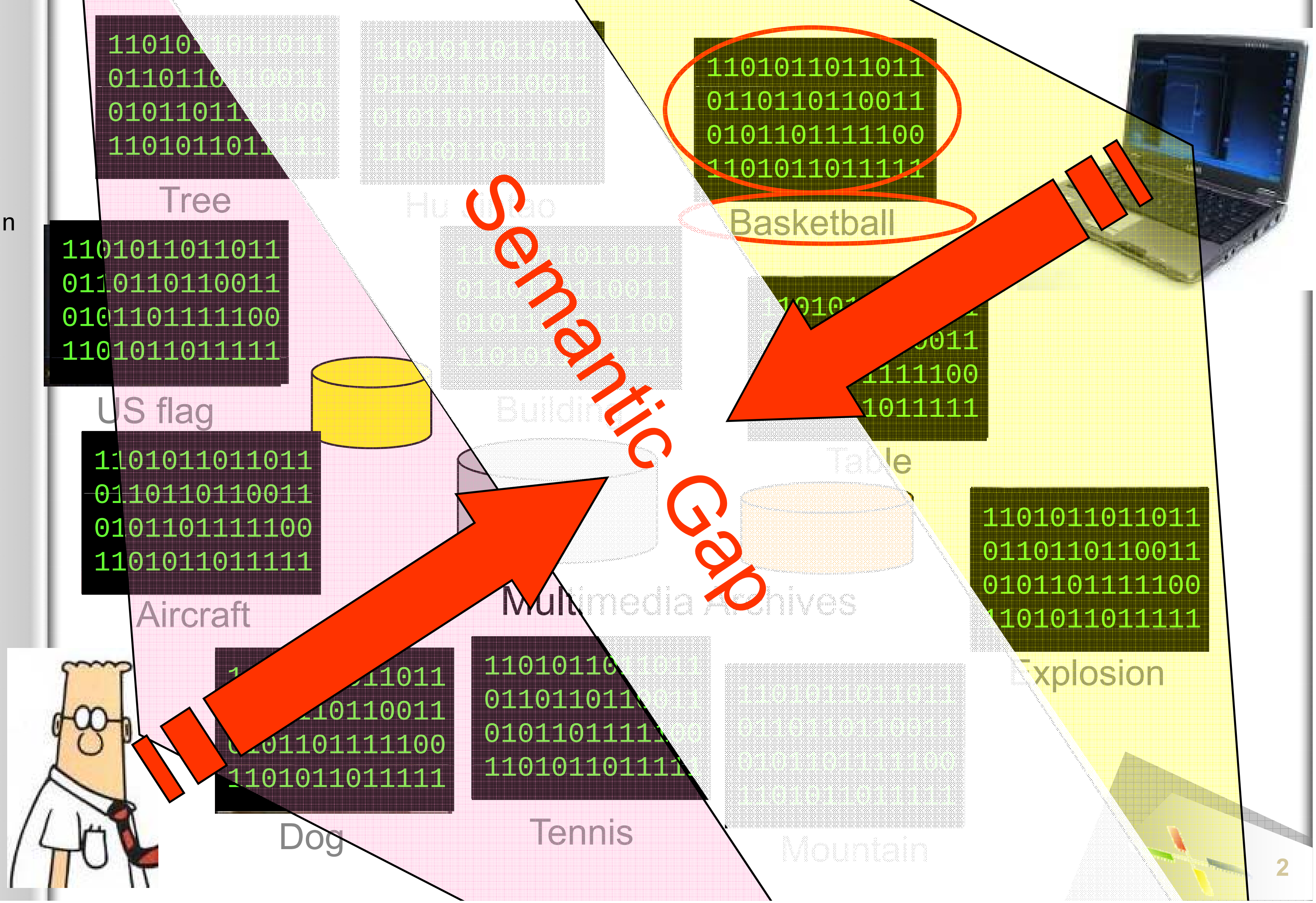
Cees Snoek

with contributions by:
many

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

Problem statement

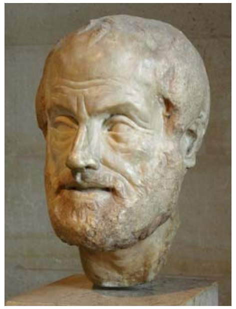
- What
- Why
- How
- Why not?
- Conclusion



- What
- Why
- How
- Why not?
- Conclusion

Where is science?

- 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

Google

naming textual information

naming rocks and minerals

naming living organisms

What about naming video information?

Societal relevance

- From broadcasting to narrowcasting

~1955

~1985

~2010

- Everybody with a message uses video to deliver it

- Growing unmanageable amounts of video

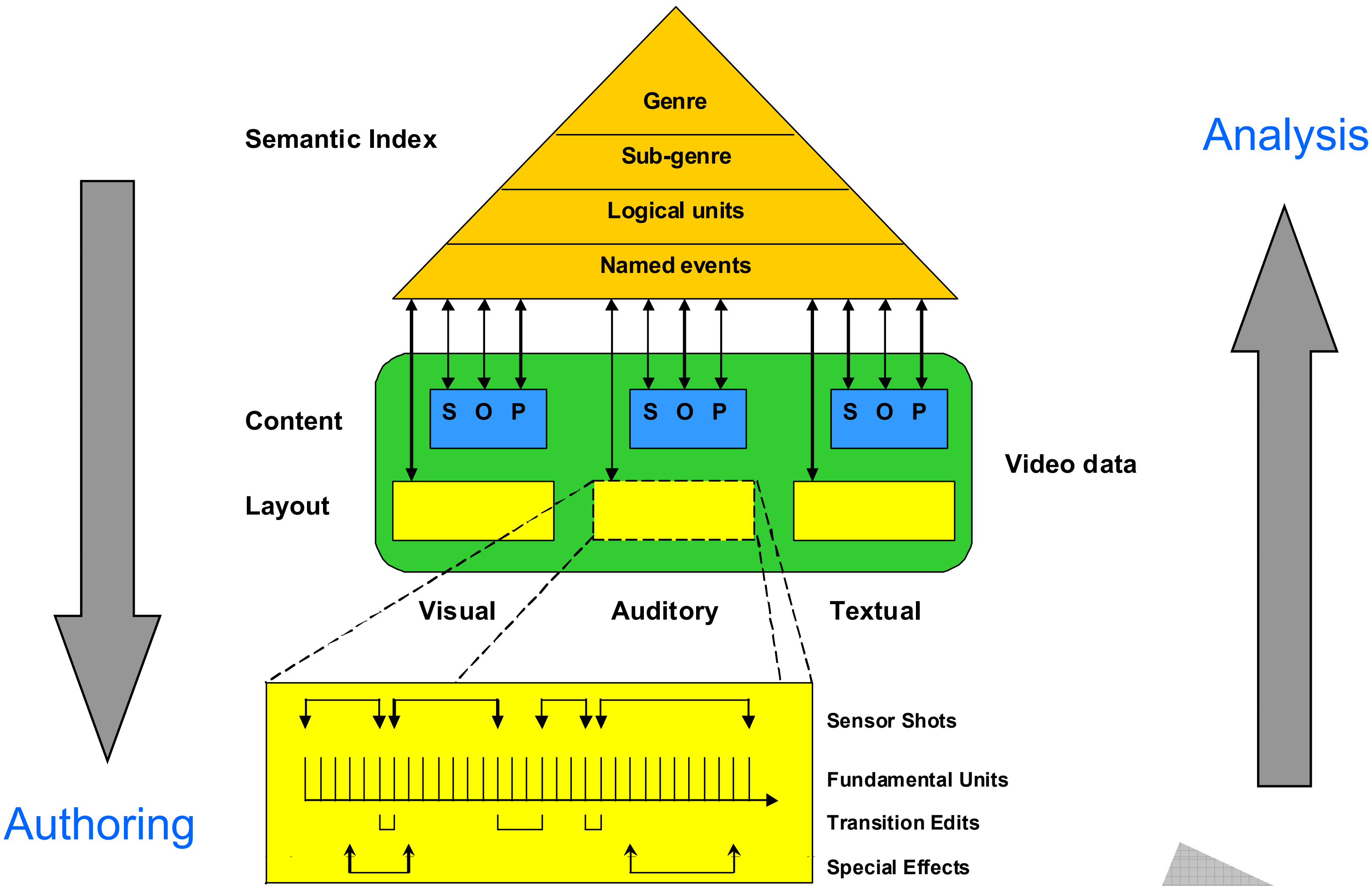
Data categories

- Produced video data
  - Definition
    - videos that are created by an author who is actively selecting content and where the author has control over the appearance of the video
  - Raw data
    - The material as it is shot
  - Edited data
    - The material that is shown in the final program
  - Digital data
    - The data as we receive it in our system

- Observed video data:
  - Definition
    - videos where a camera is recording some scene and where the author does not have the means to manipulate or plan the content.

Author's framework

- What
- Why
- How
- Why not?
- Conclusion



The goal: semantic video indexing

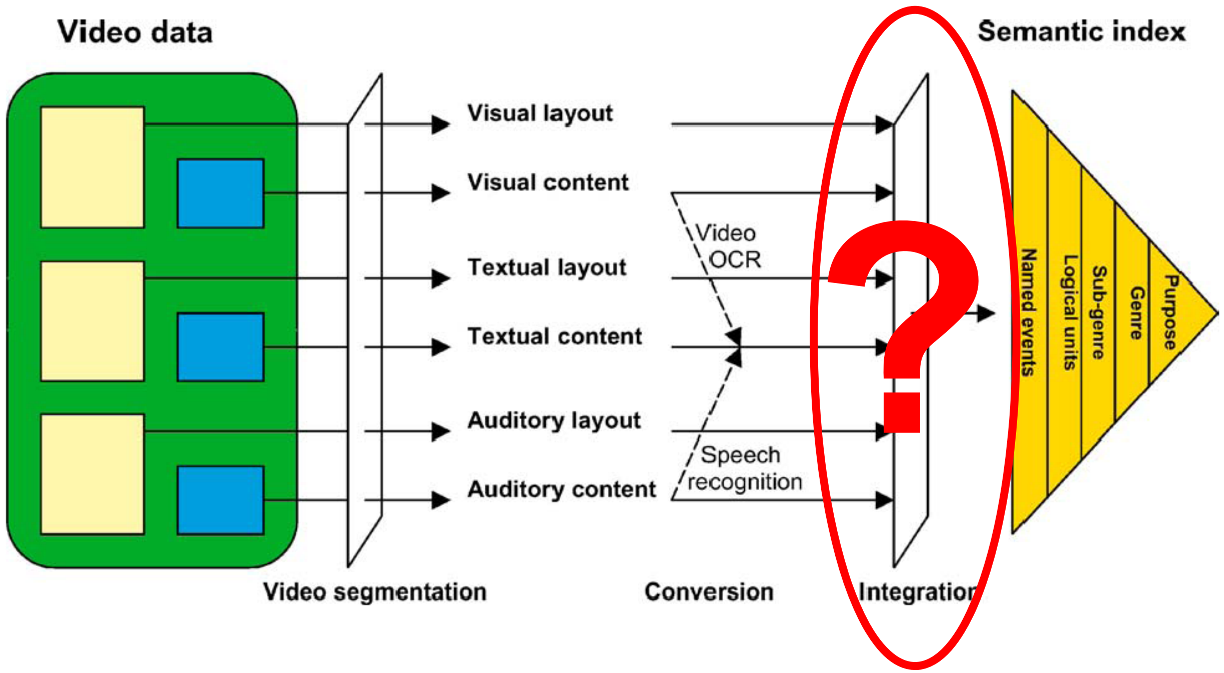
- What
- Why
- How
- Why not?
- Conclusion

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

Airplane

Authoring-driven analysis

- What
- Why
- How
- Why not?
- Conclusion



➢ How to obtain a reliable semantic index?

Semantic indexing

- What
- Why
- How
- Why not?
- Conclusion

➢ 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

So how about these?

- What
- Why
- How
- Why not?
- Conclusion


Animal


Building


Road


Beach


Boat

Graphic


People


Car


Vegetation

Overlayed Text


Studio Setting


Outdoor


Train

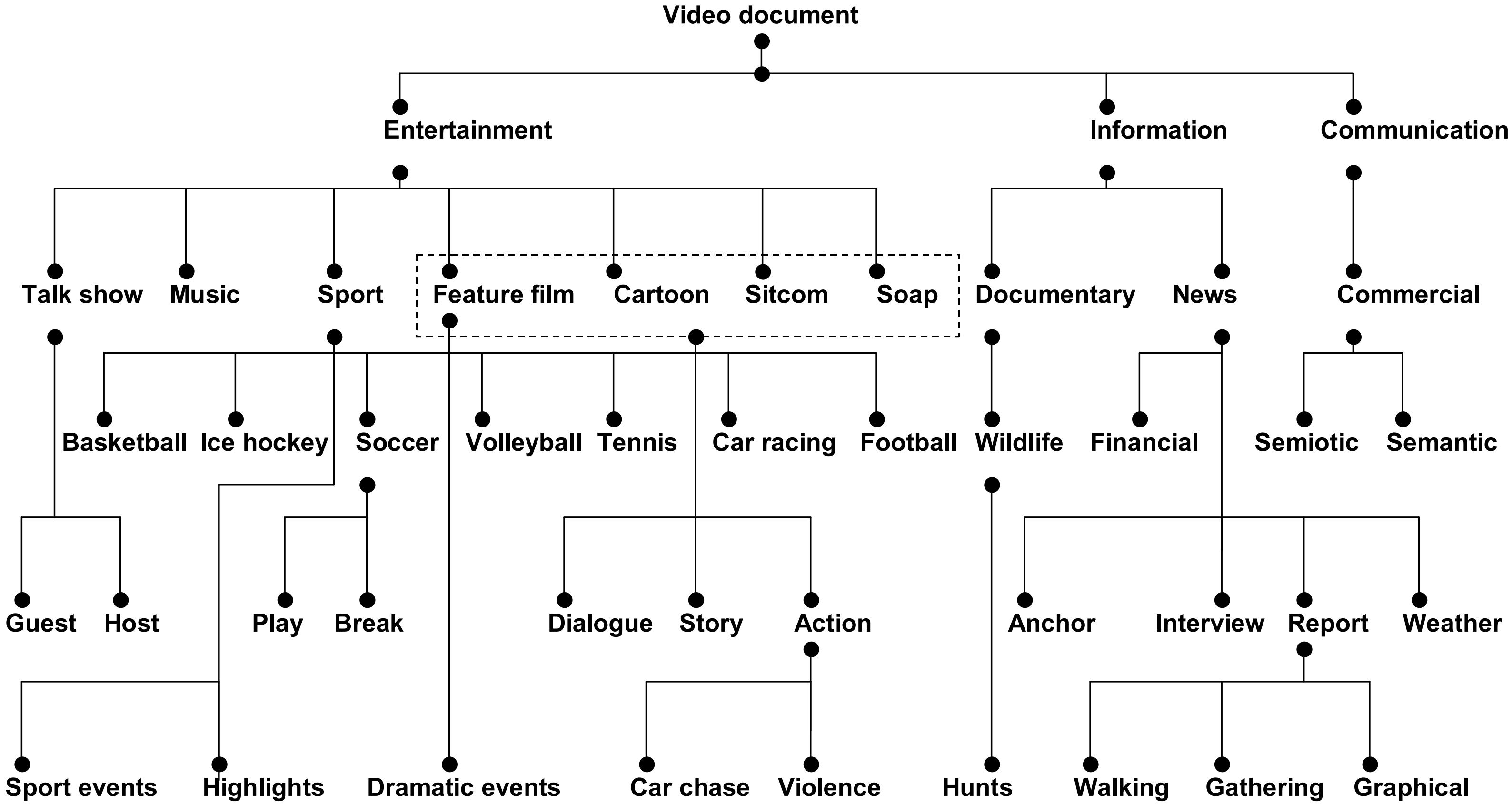

Bicycle

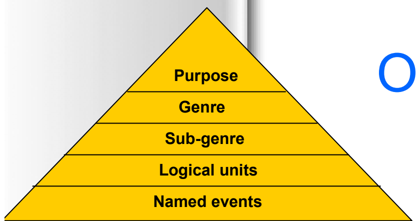
And the > 1000 others ………

Semantic index overview

+/- 2001

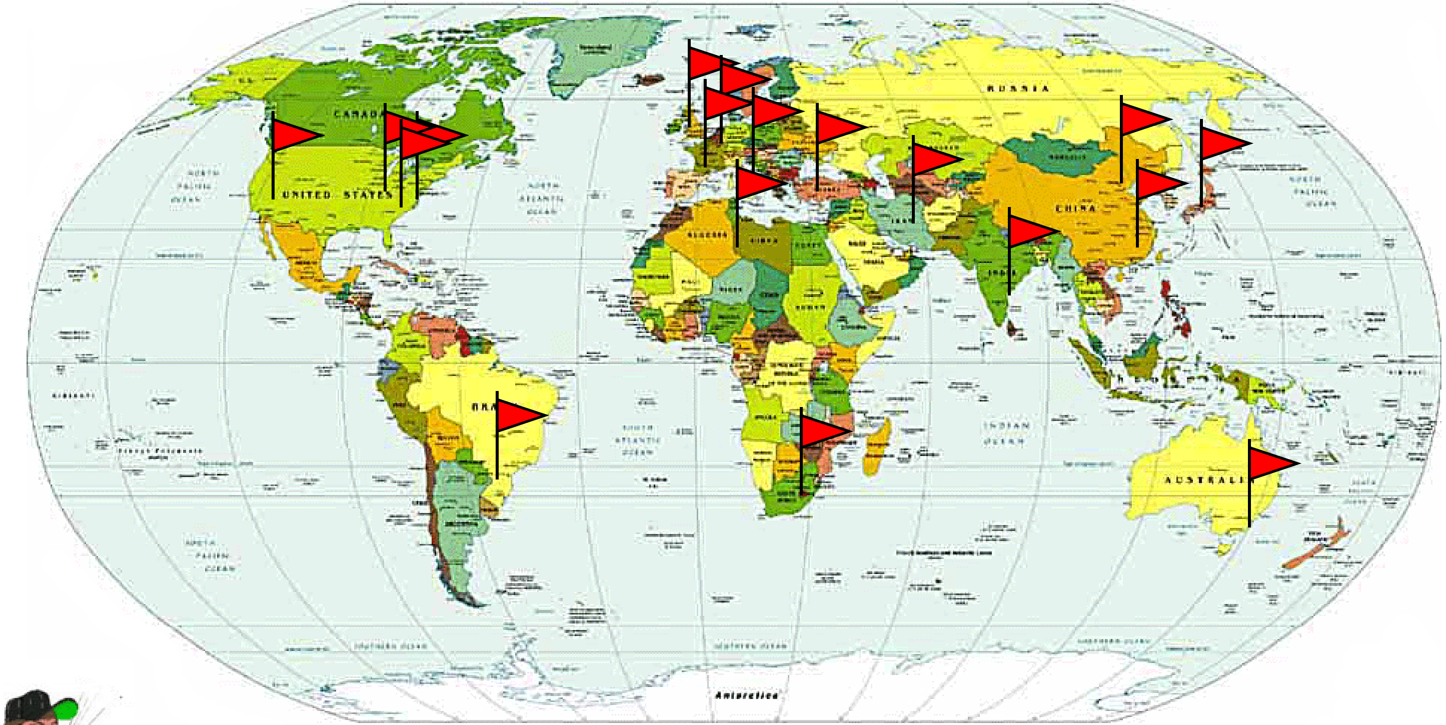
- What
- Why
- How
- Why not?
- Conclusion





One PhD per detector requires too many students…

Fragmented research efforts...



NIST

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 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

abc

Ground truth

Data set

Speech transcript

- Large international field of participants

Carnegie Mellon · FXPAL · 清华大学 Tsinghua University · University of Oulu · NUS School of Computing · IBM

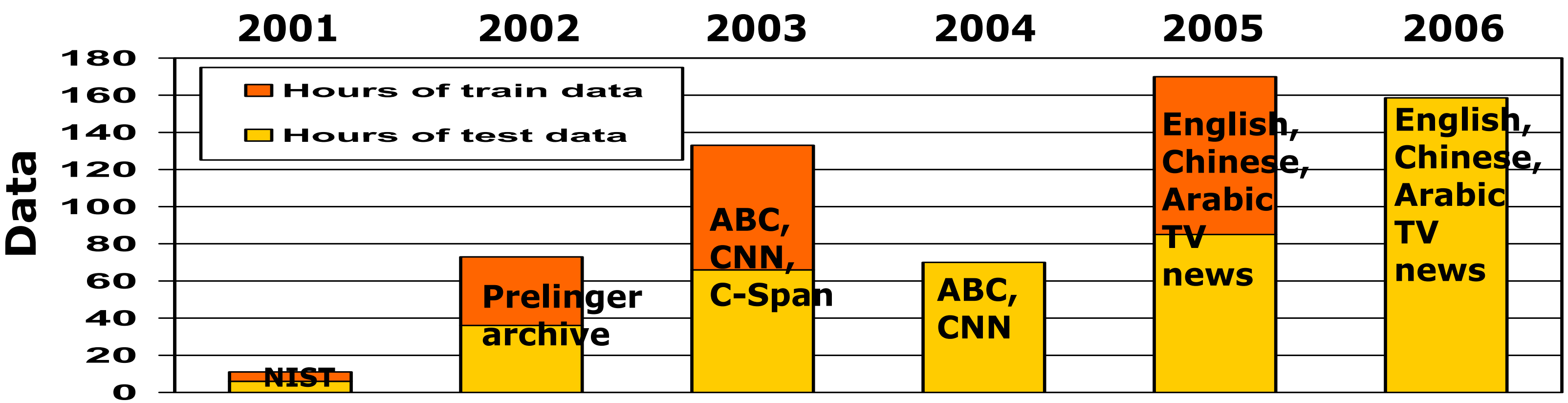
- Currently the de facto standard for evaluation

http://trecvid.nist.gov/

Source: Paul Over, NIST

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

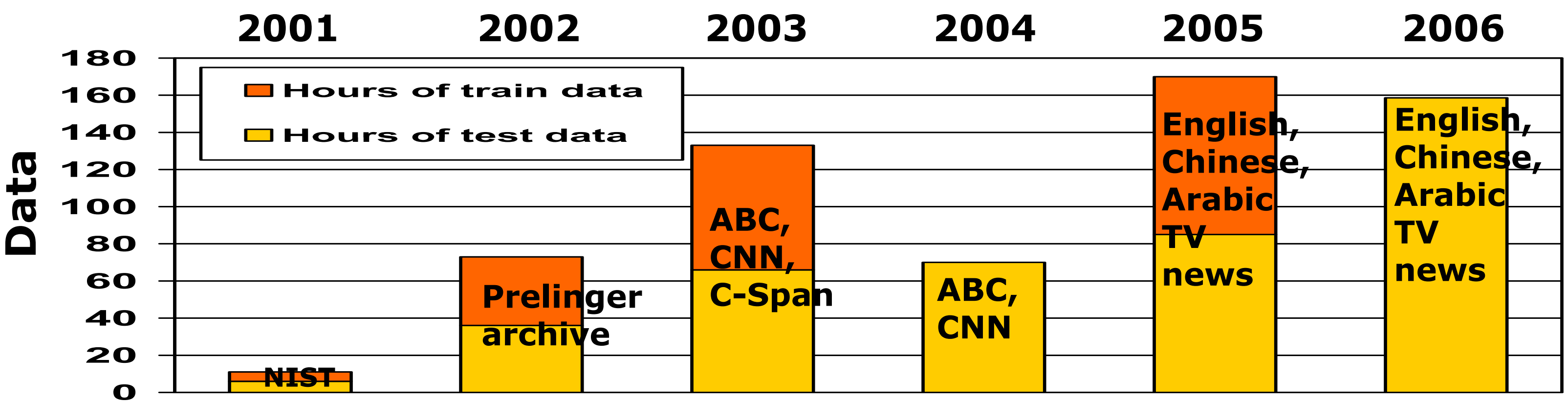
- What
- Why
- How
- Why not?
- Conclusion



National Institute of Standards and Technology

Source: Paul Over, NIST

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



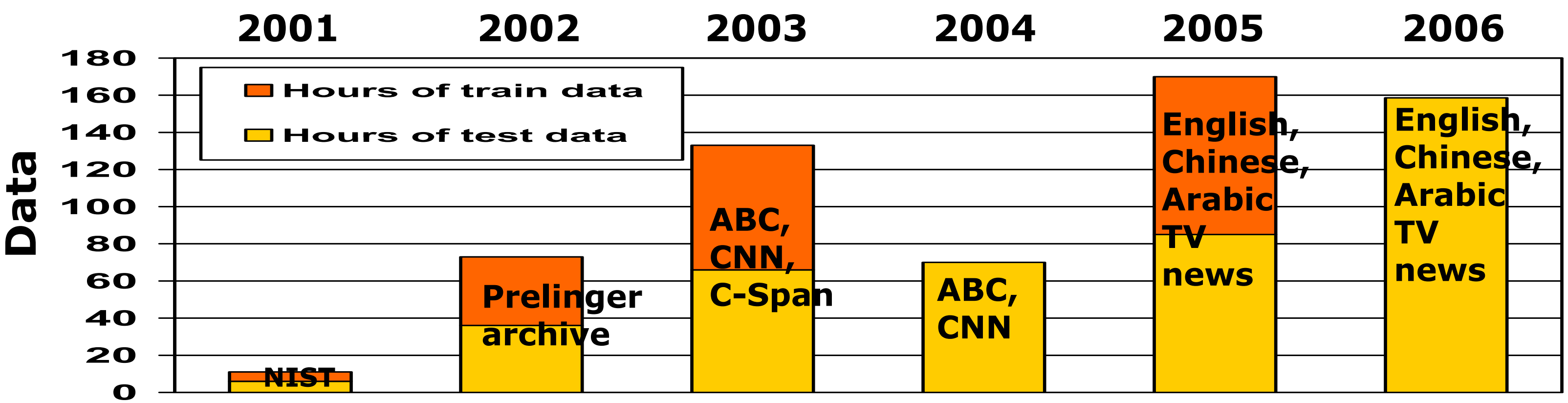
Tasks

2001	2002	2003	2004	2005	2006
Shots	Shots	Shots	Shots	Shots	Shots
Search	Search	Search	Search	Search	Search
	Concepts	Concepts	Concepts	Concepts	Concepts
		Stories	Stories	BBC rushes	BBC rushes
				Camera motion	

Source: Paul Over, NIST

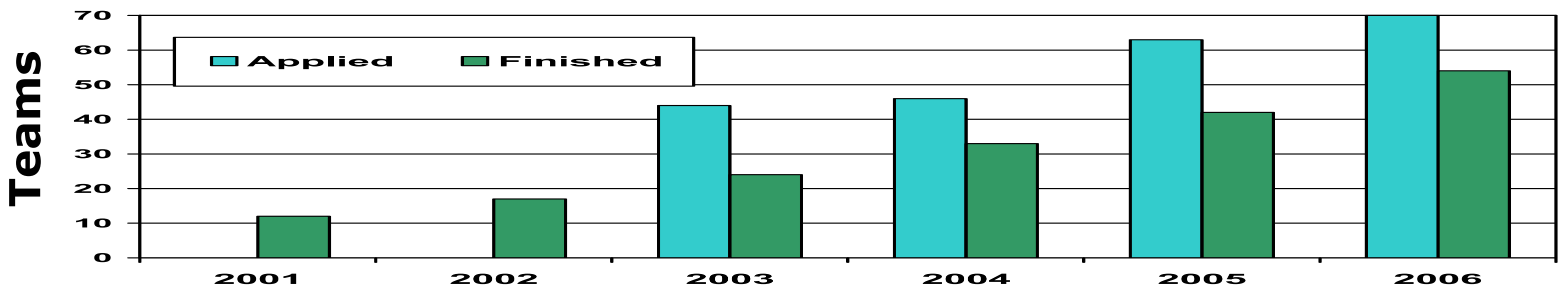
TRECVID Evolution: data, tasks, participants,...

- What
- Why
- How
- Why not?
- Conclusion



Tasks

2001	2002	2003	2004	2005	2006
Shots	Shots	Shots	Shots	Shots	Shots
Search	Search	Search	Search	Search	Search
	Concepts	Concepts	Concepts	Concepts	Concepts
		Stories	Stories	BBC rushes	BBC rushes
				Camera motion	

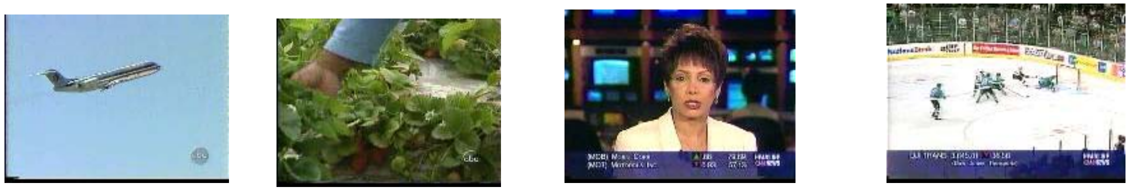


Peer-reviewed papers:	2001	2002	2003	2004	2005	2006
		10	18	51	40	55

NIST
National Institute of Standards and Technology

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*

TRECVID evaluation measures

- What
- Why
- How
- Why not?
- Conclusion

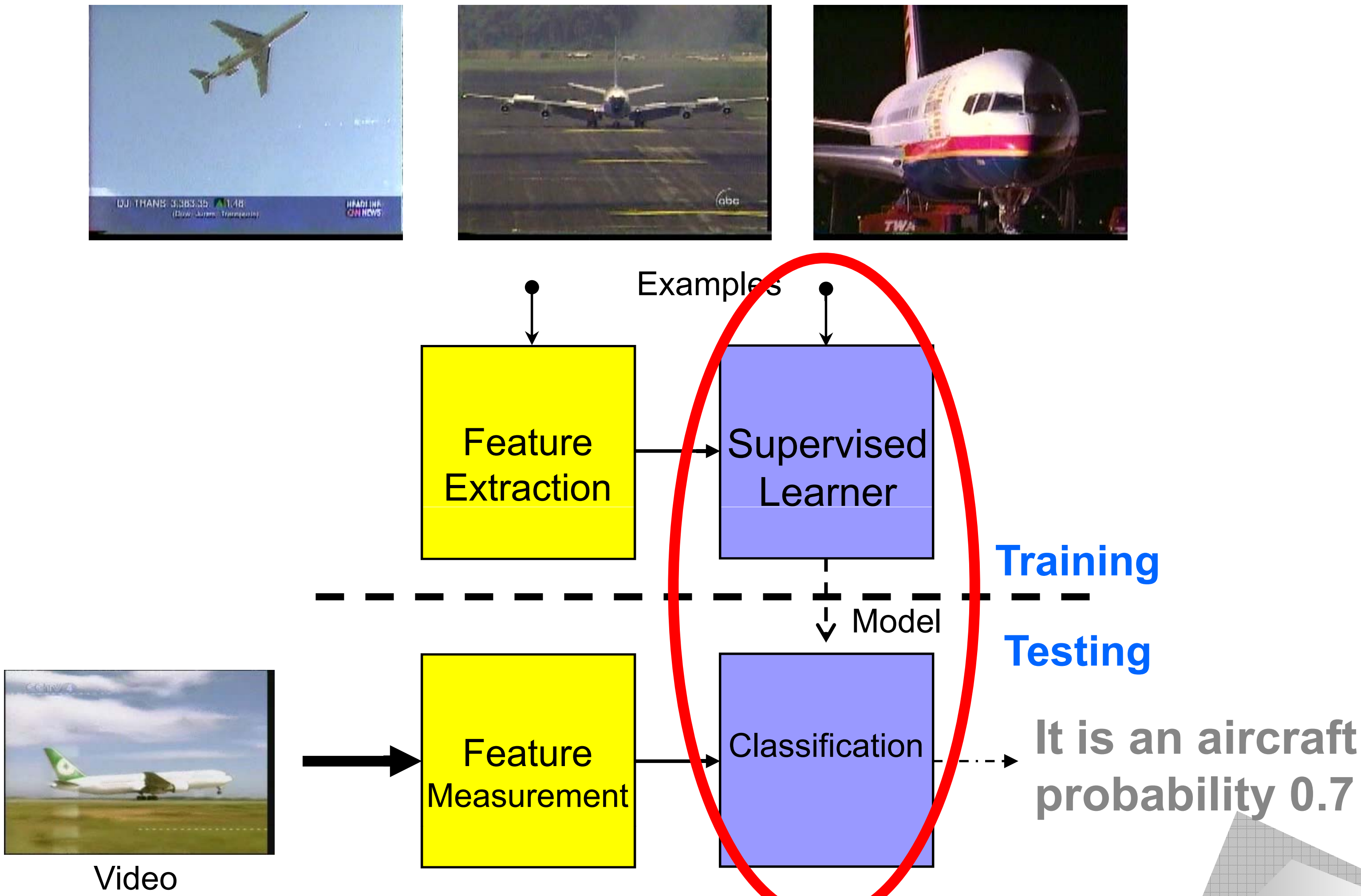
- Classification procedure
  - ✓ Training: many hours of (partly) annotated video
  - ✓ Testing: many hours of unseen video

- Evaluation measure: Average Precision
  - ✓ Combines precision and recall
  - ✓ Averages precision after every relevant shot
  - ✓ Top of the ranked list most important

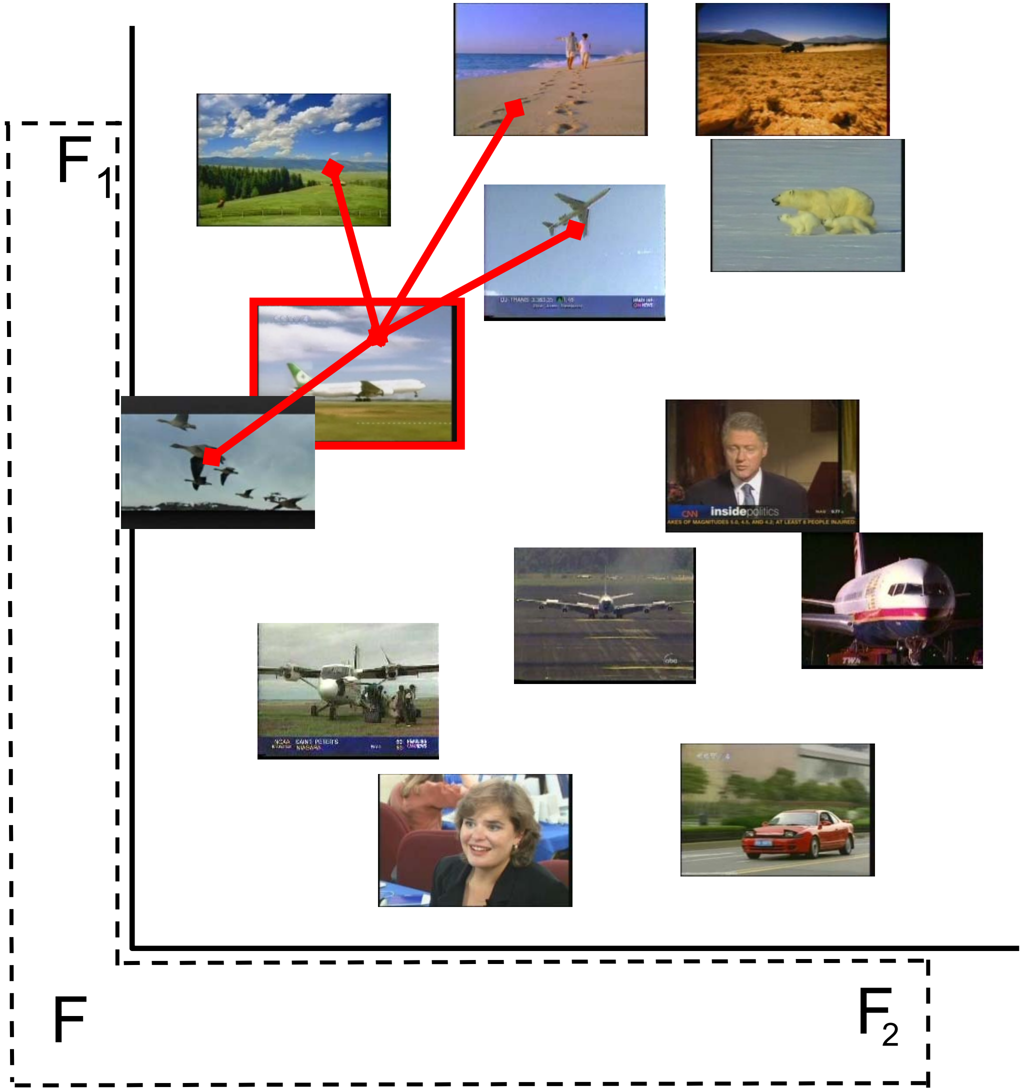
$$AP = \frac{1/1 + 2/3 + 3/4 + \ldots}{\text{Total Number of correct shots}}$$

A simple concept detector

- What
- Why
- How
- Why not?
- Conclusion

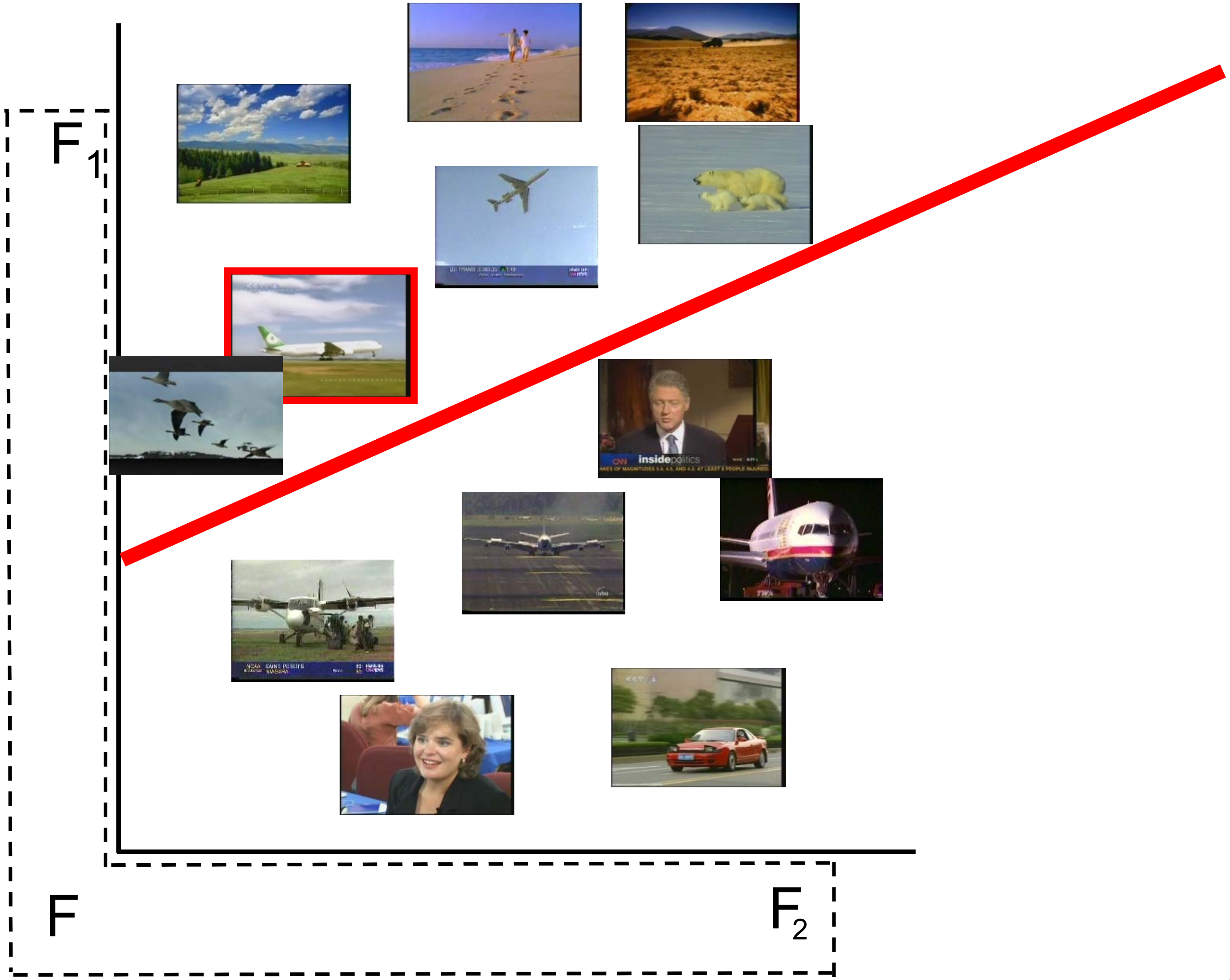


K nearest neighbor



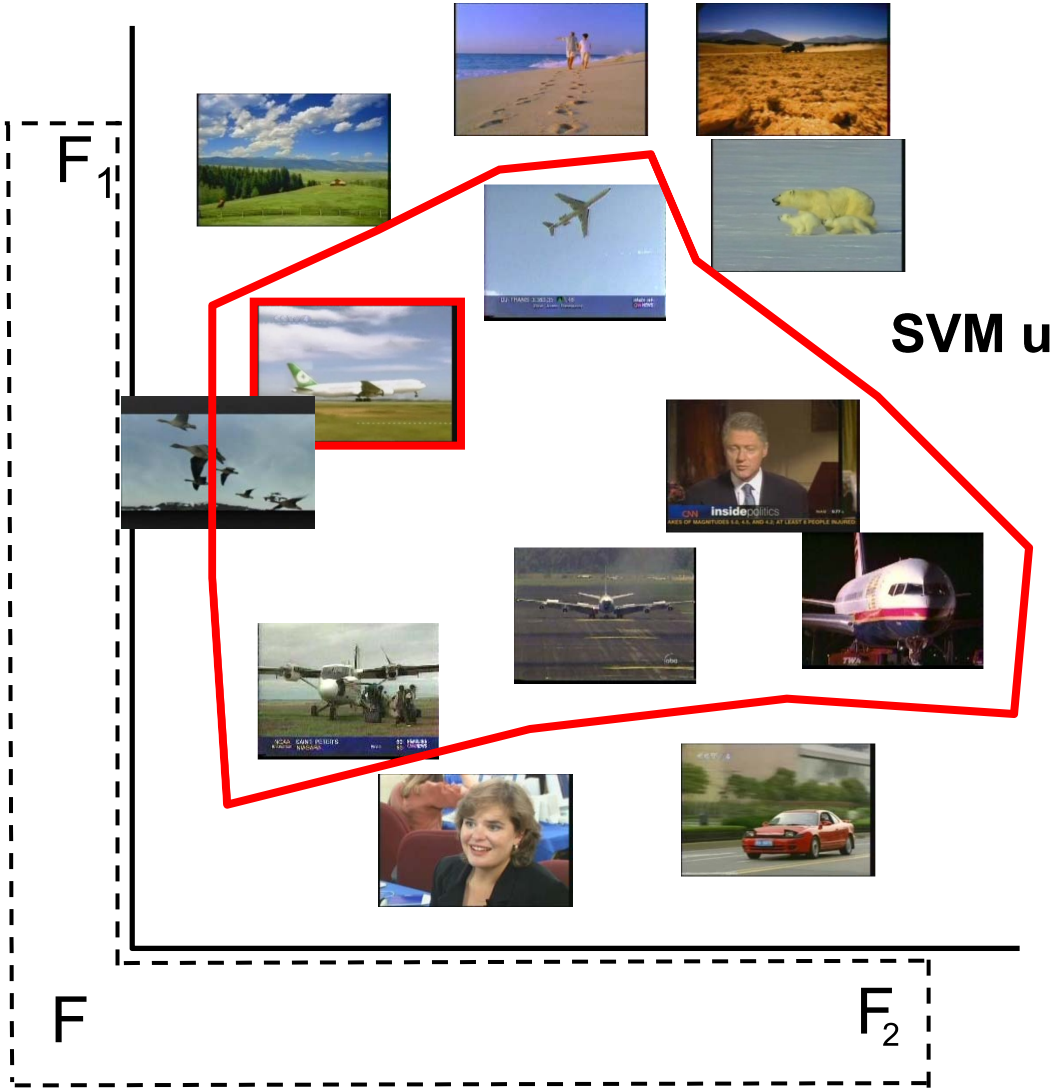
Linear classification

- What
- Why
- How
- Why not?
- Conclusion



Support vector machine

- What
- Why
- How
- Why not?
- Conclusion



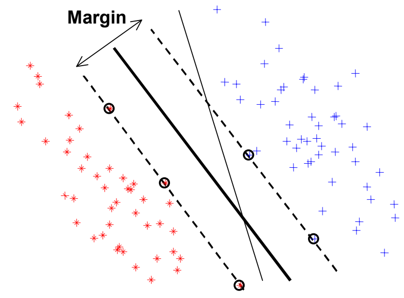
SVM usually is a good choice

- What
- Why
- How
- Why not?
- Conclusion

Supervised Learner

➢ Support Vector Machine

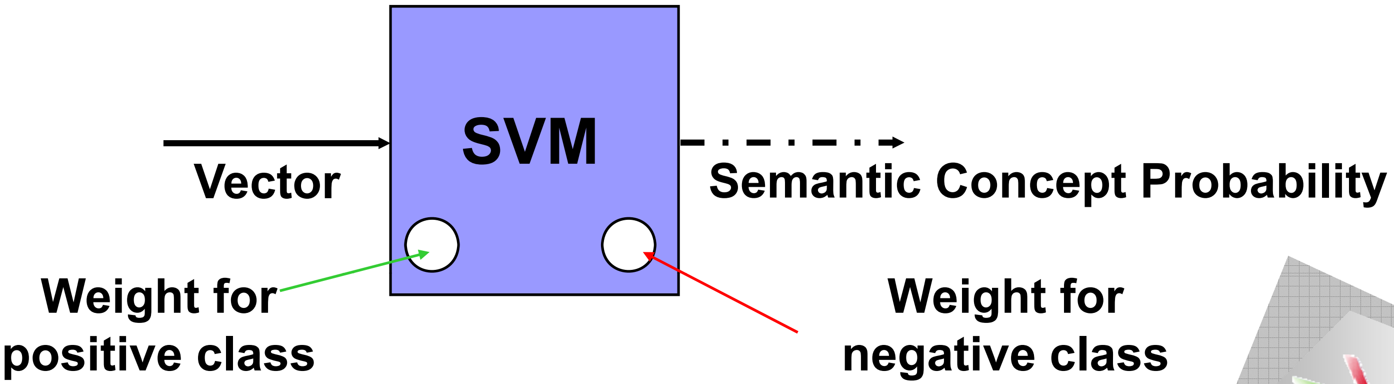
- ✓ Learns from provided examples
- ✓ Maximizes margin between two classes
- ✓ Problematic when data not balanced



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

➢ Solution for balancing problem

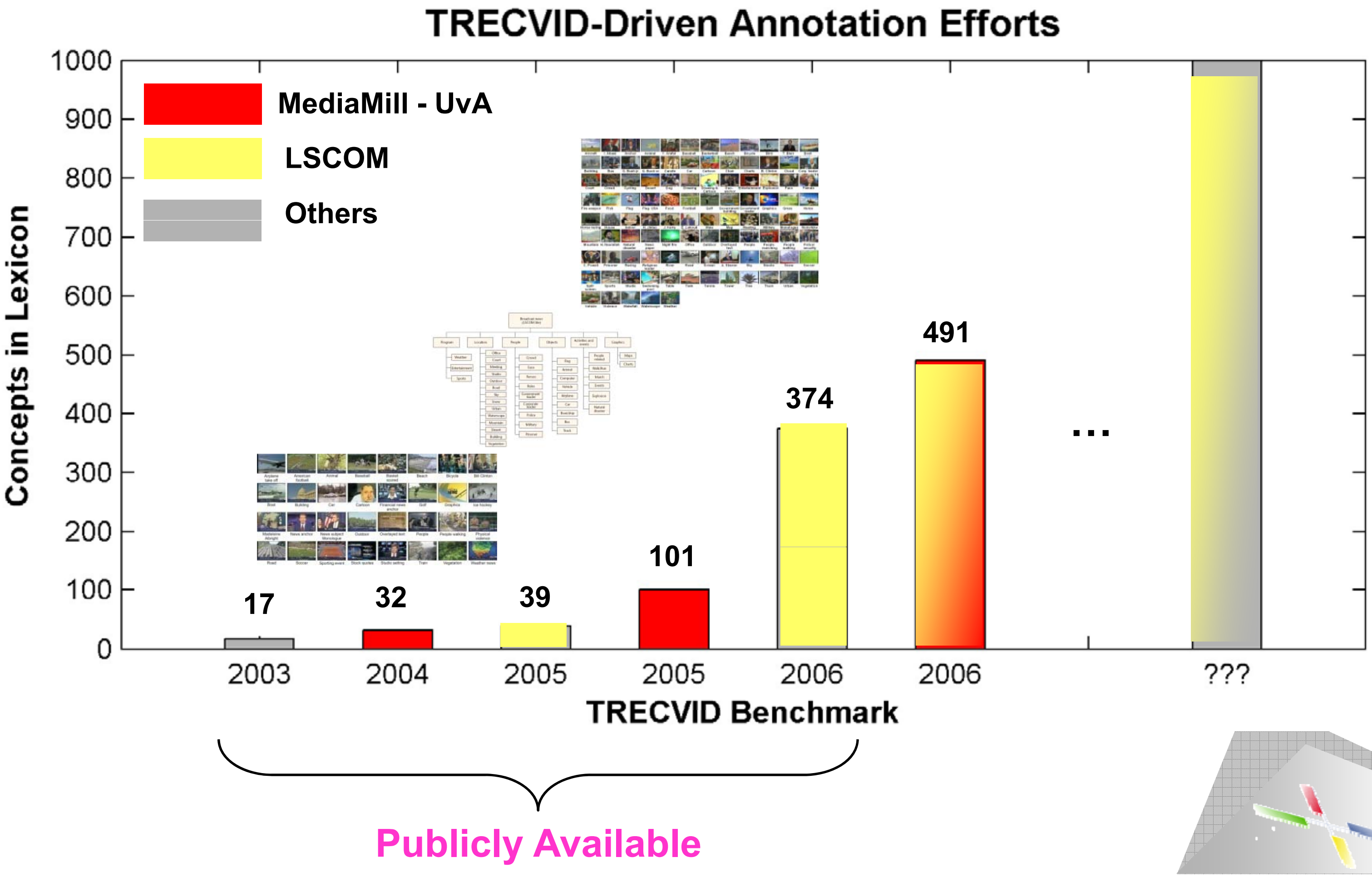
- ✓ Adapt penalty parameters in SVM formulation
- ✓ Select best of multiple weight combinations
- ✓ Using cross validation (even more examples needed)



- What
- Why
- How
- Why not?
- Conclusion

Concept detector: requires examples

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



Concept definition

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

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

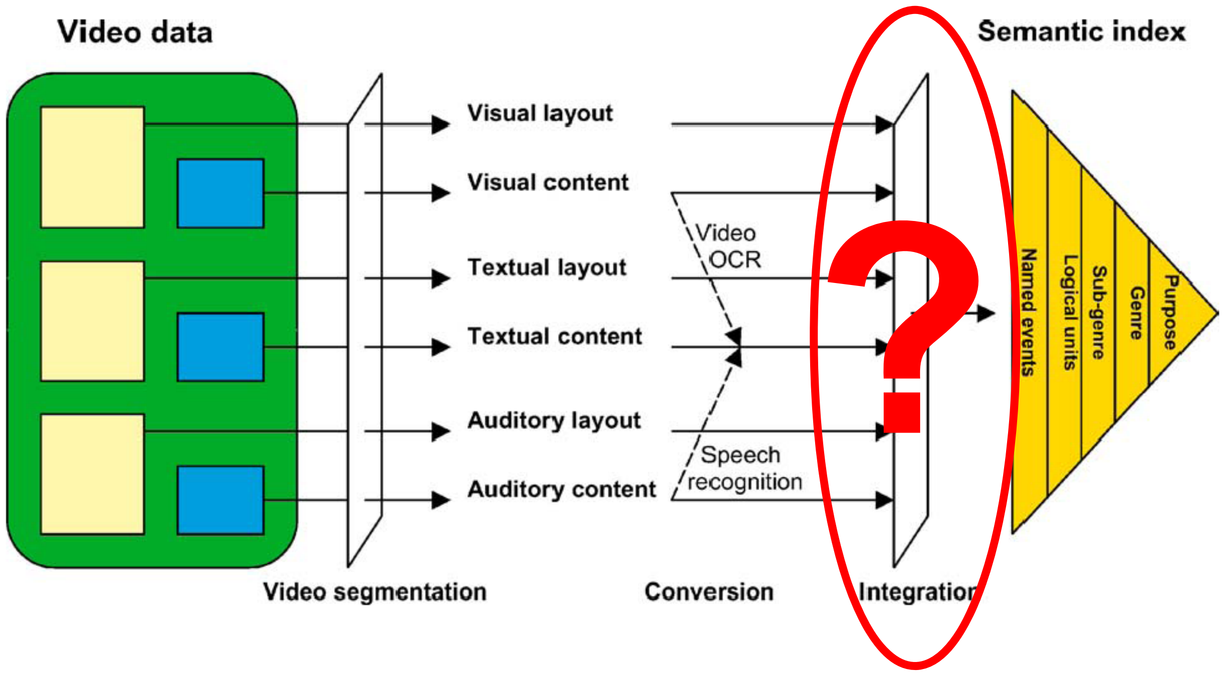
Collaborative annotation tool

TRECVID 2005

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

Recap: authoring-driven analysis

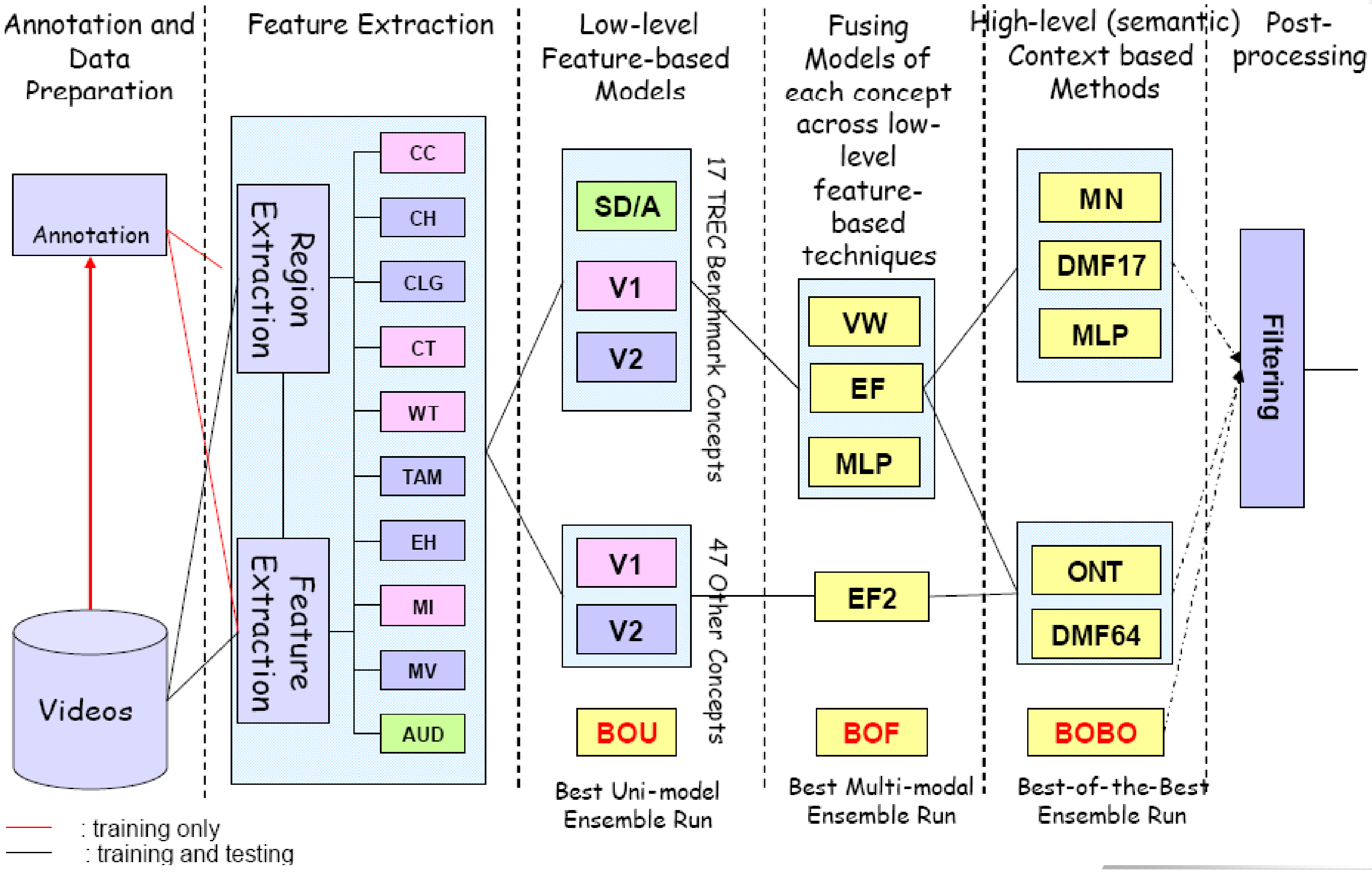
- What
- Why
- How
- Why not?
- Conclusion



➢ How to obtain a reliable semantic index?

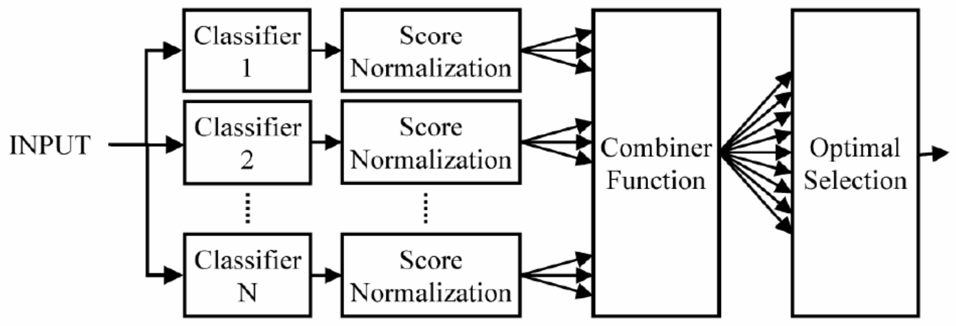
IBM's pipeline

- What
- Why
- How
- Why not?
- Conclusion



IBM’s pipeline approach

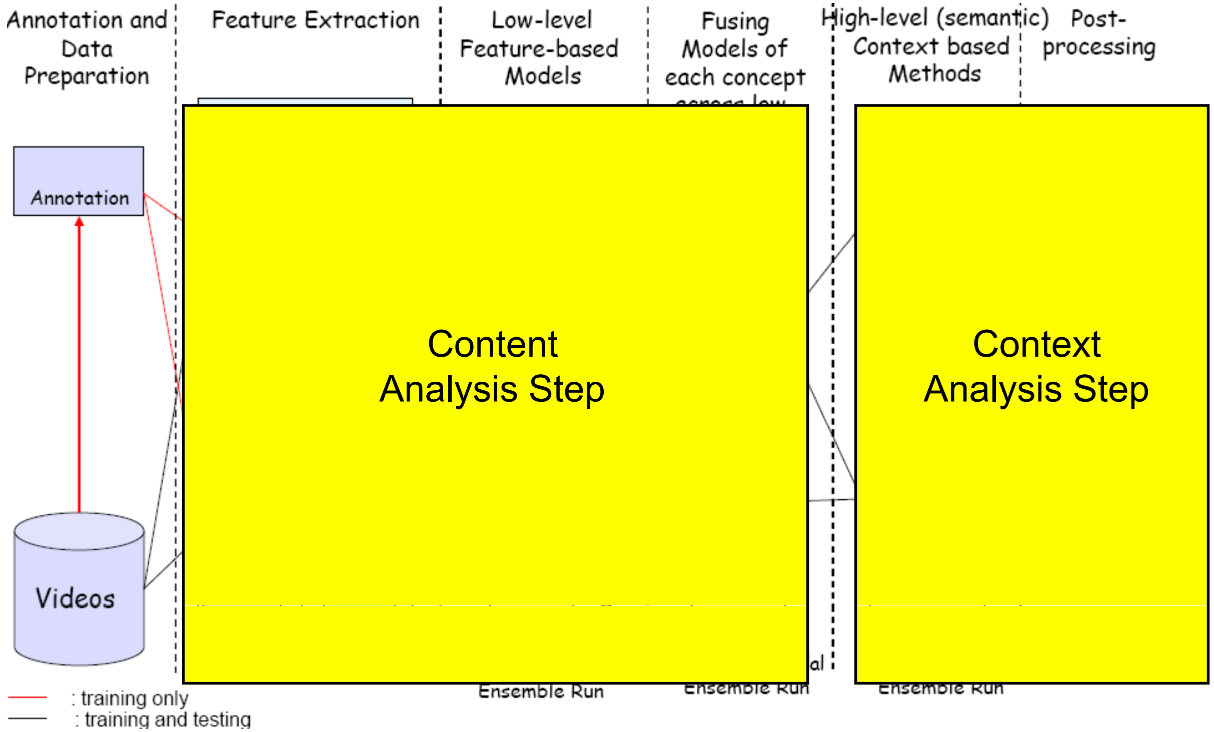
- Split train data into several sets
  - ✓ Each pipeline has different validation set
- Optimize all classifier configurations
  - ✓ Set of basic image, audio, and text features
  - ✓ Set of unimodal models for lexicon of concepts
- Experiment with different fusion methods
  - ✓ SVM, NN, Multinet, ensembles, ontologies,…



IBM's pipeline approach

TRECVID lessons

- What
- Why
- How
- Why not?
- Conclusion

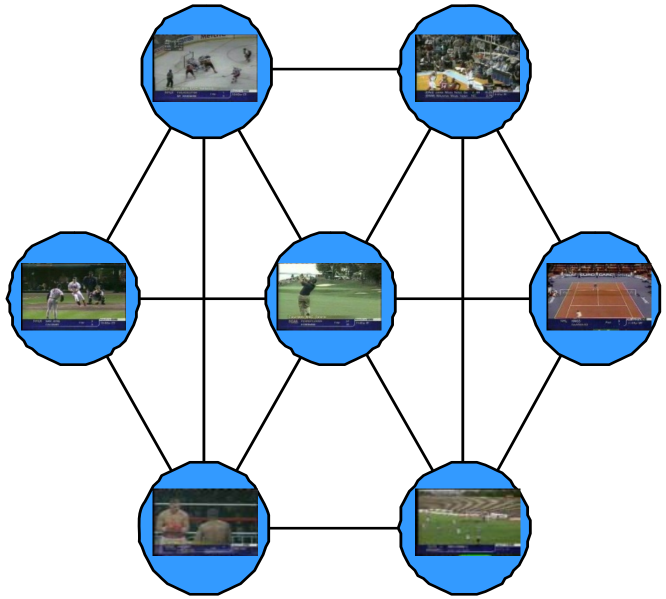


- First generic video indexing approach
  - ✓ Highly successful in TRECVID benchmark
  - ✓ Combines machine learning with content and context abstractions

Context

TRECVID lessons

- Exploitation of context for video analysis
  - ✓ Concepts do not occur in vacuum
  - ✓ In contrast, they are interconnected

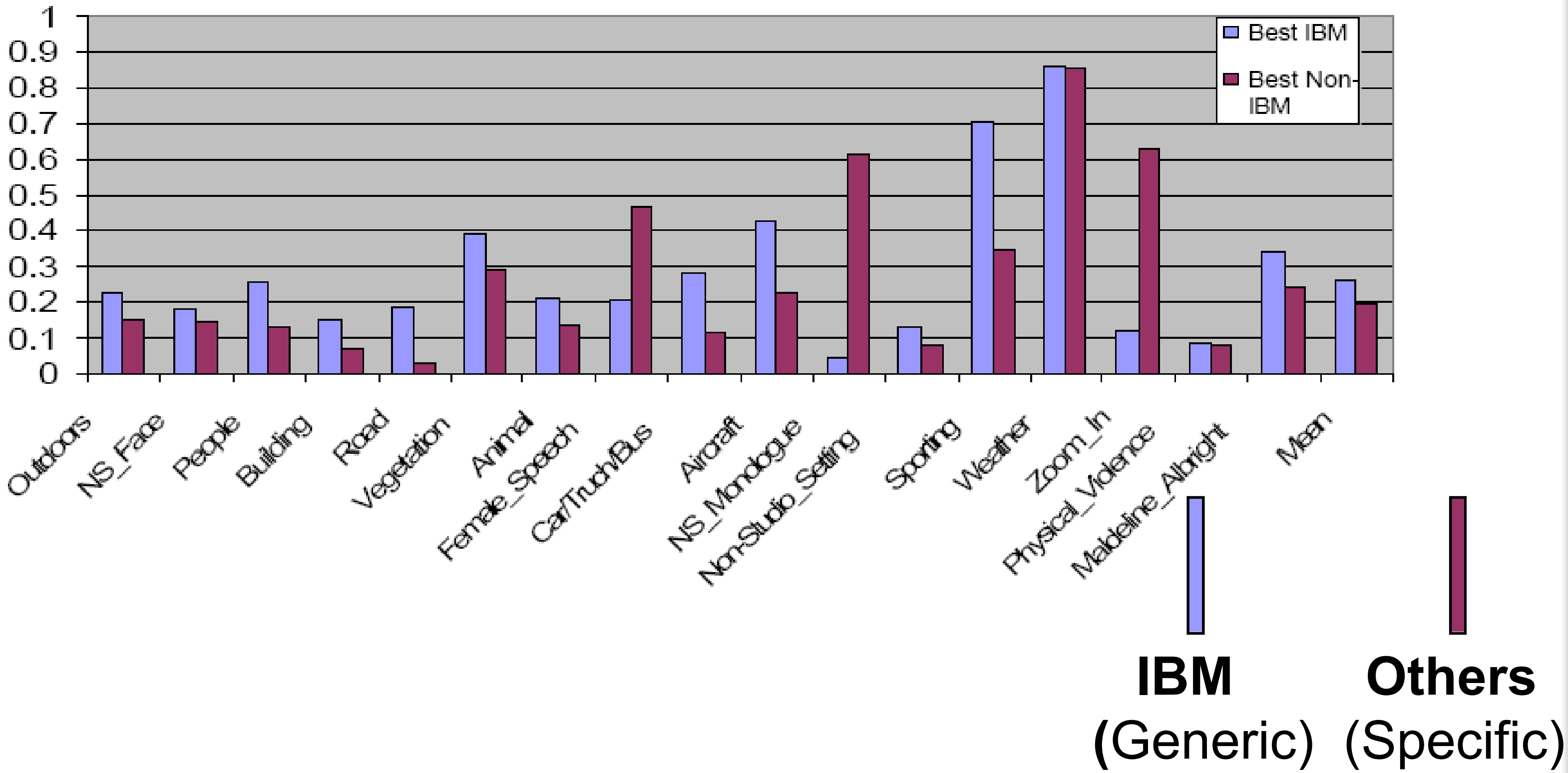


- What is sports?
  - ✓ Answer: a combination of various individual sports

Concept detection task

- What
- Why
- How
- Why not?
- Conclusion

TRECVID 2003

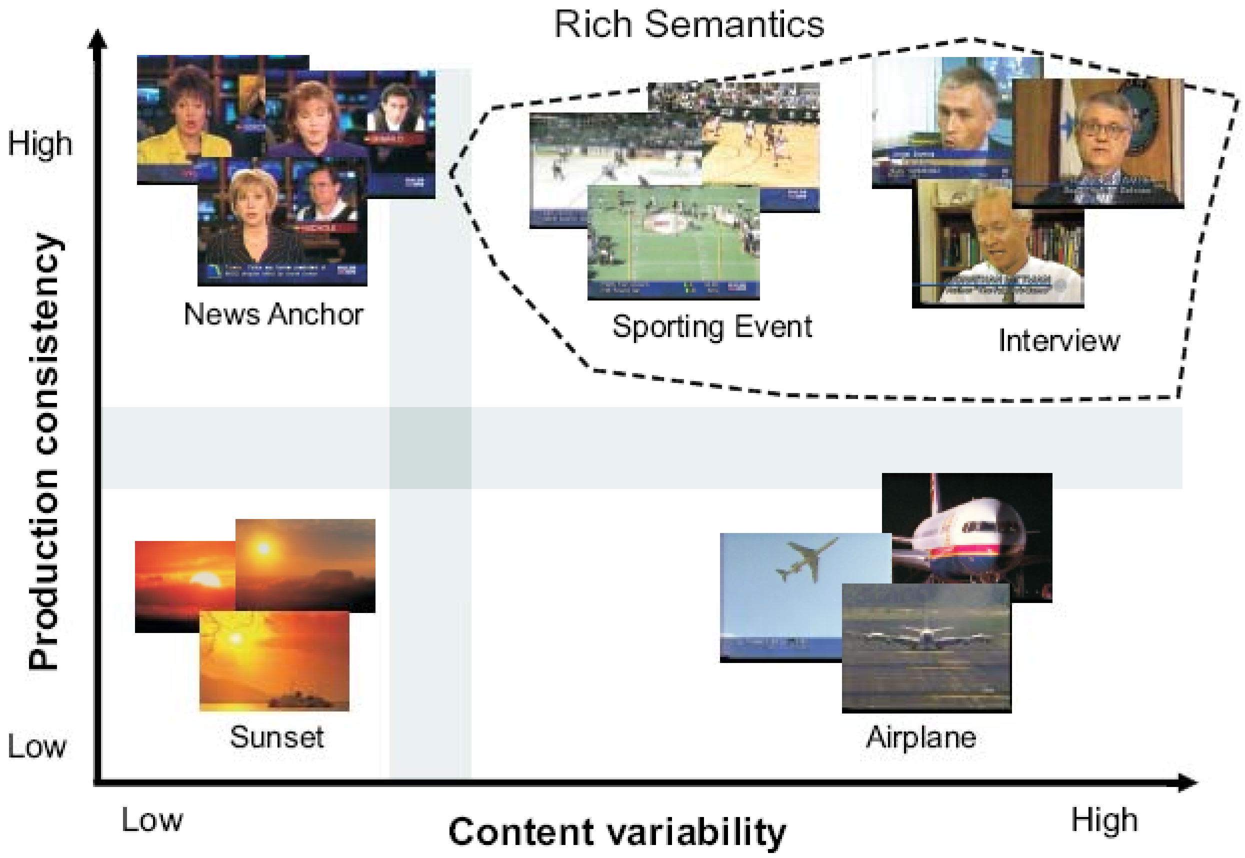


- IBM's approach works very well on (visual) concepts related to content
- How about rich semantic concepts?
  - ✓ With large variability in production process

Rich semantics

TRECVID lessons

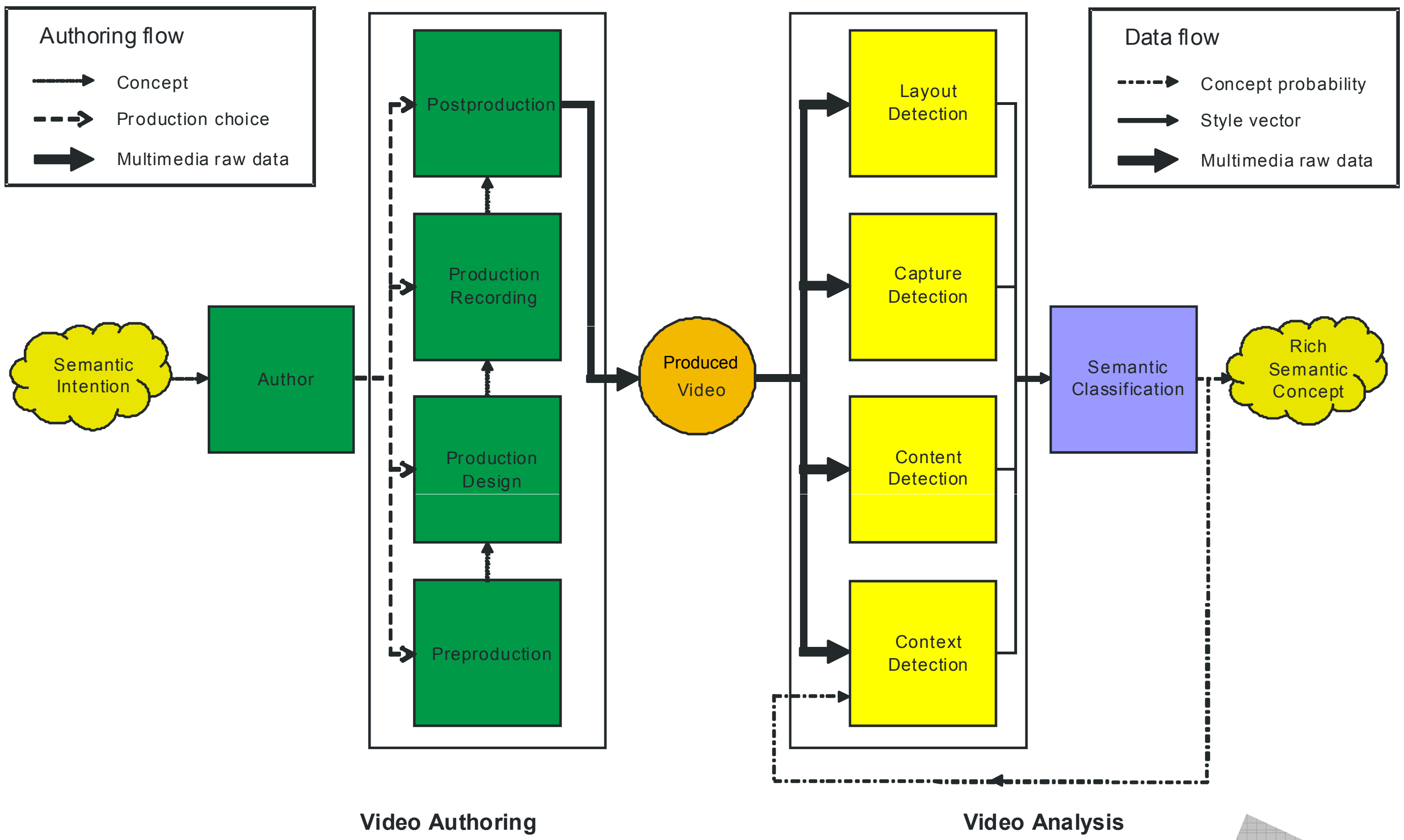
- What
- Why
- How
- Why not?
- Conclusion



Production style

TRECVID lessons

- What
- Why
- How
- Why not?
- Conclusion

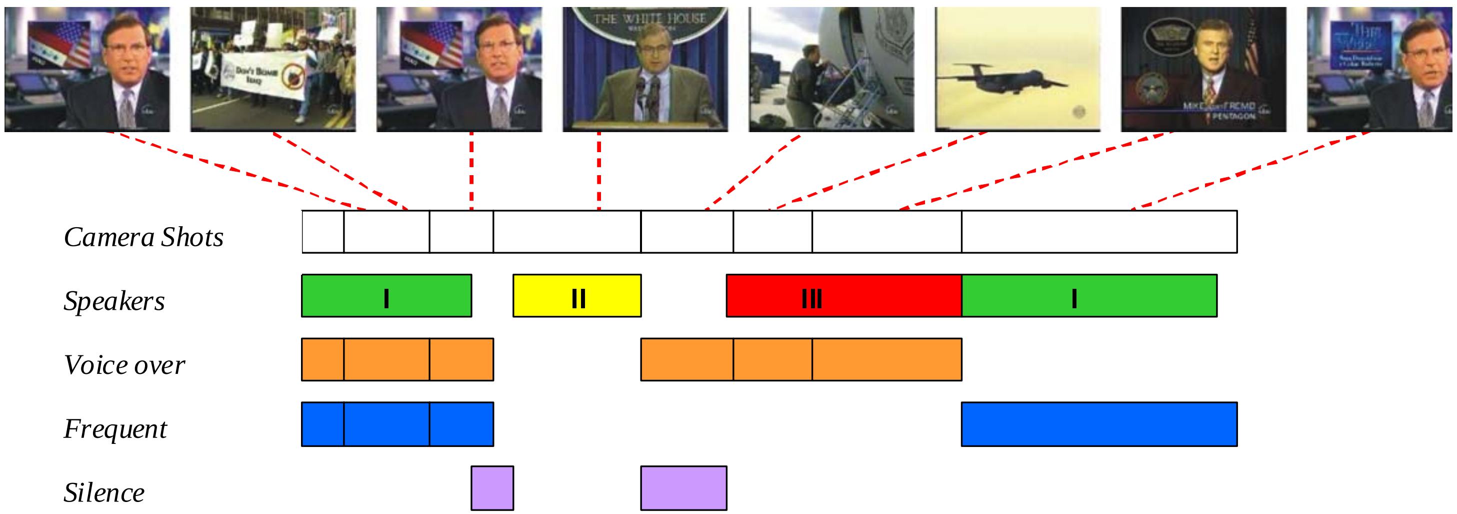


References:
Gauvain et al, Sp. Comm. '02
Informedia, CMU

Style detectors

TRECVID lessons

- What
- Why
- How
- Why not?
- Conclusion



References:
Gauvain et al, Sp. Comm. '02
Informedia, CMU

Style detectors

- What
- Why
- How
- Why not?
- Conclusion

TRECVID lessons

"Ann Compton
ABC news,
New York"

Jemmy Curter

match?

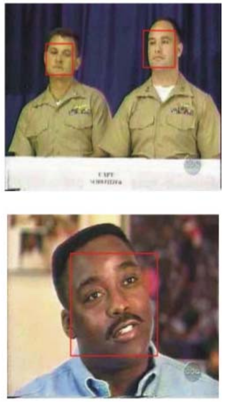
isName?

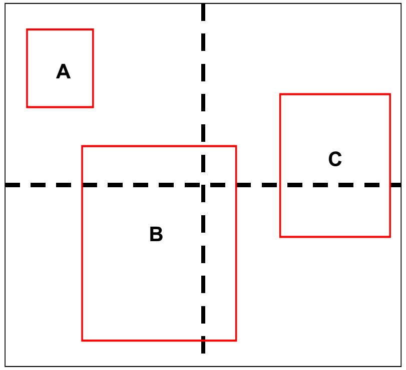
"Call now:
1800…"

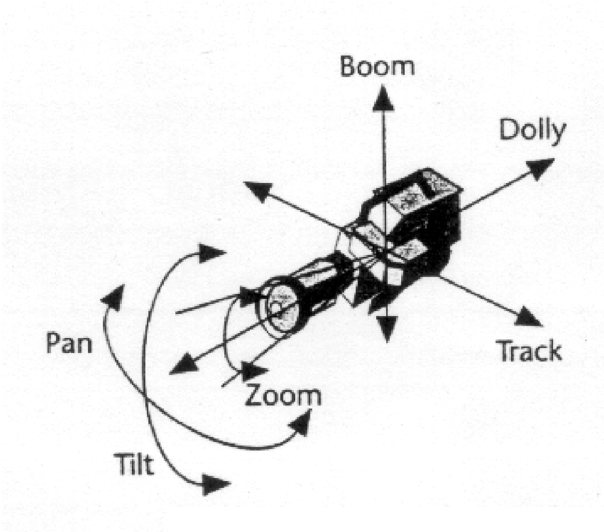
References:
Schneiderman et al, IJCV '04
Informedia, CMU

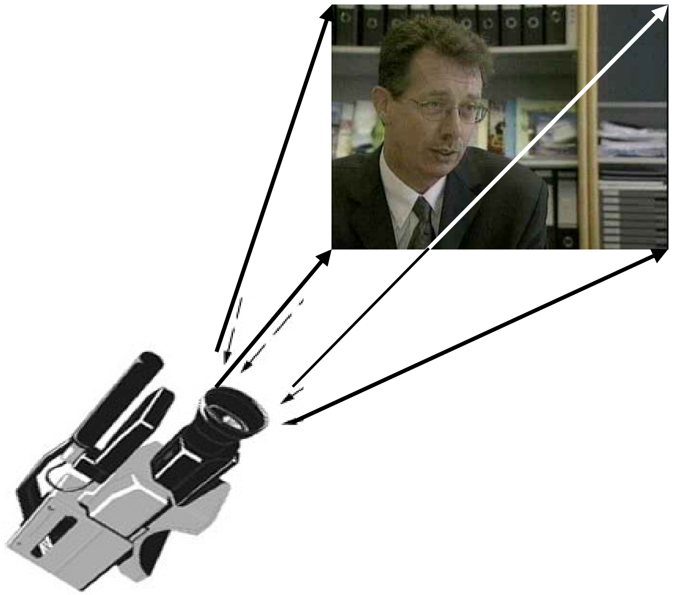
Style detectors

TRECVID lessons









Concept detection task

TRECVID 2003

- What
- Why
- How
- Why not?
- Conclusion

Key frame based analysis

TRECVID lessons

Shot

Key Frame

Visual Analysis

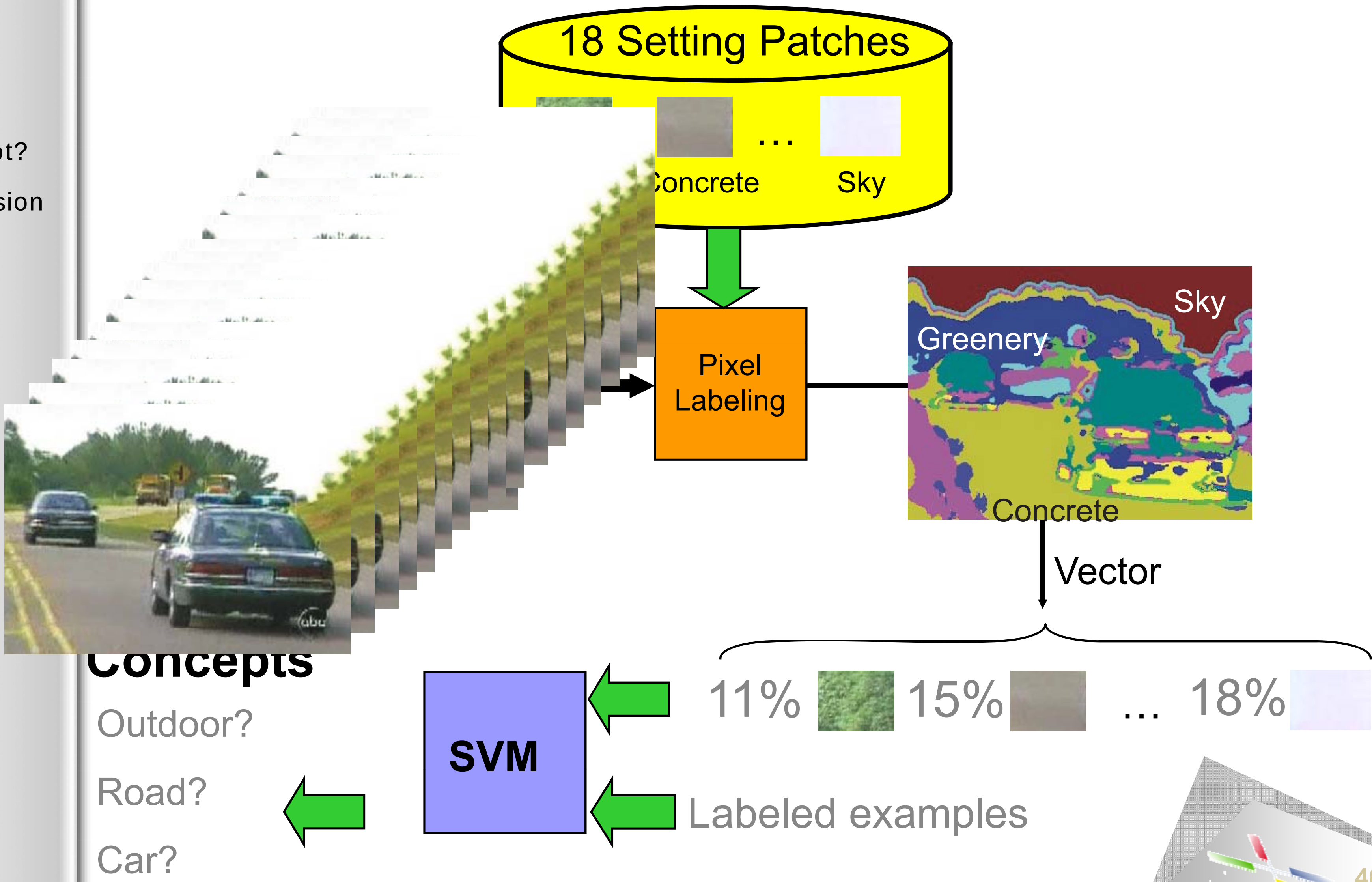
+ OK when content is static
- Not OK when content changes
- Not OK when shot segmentation is imperfect

Need for analysis beyond the key frame?

A visual analysis scenario

TRECVID lessons

- What
- Why
- How
- Why not?
- Conclusion



Frame combination functions

TRECVID lessons

Shot

Frame 1 Frame 2 Frame n

$P(car) = 0.81$ $P(car) = 0.78$ … $P(car) = 0.77$

Shot-based combination: $1/n \, \Sigma \, P(car \mid frame)$

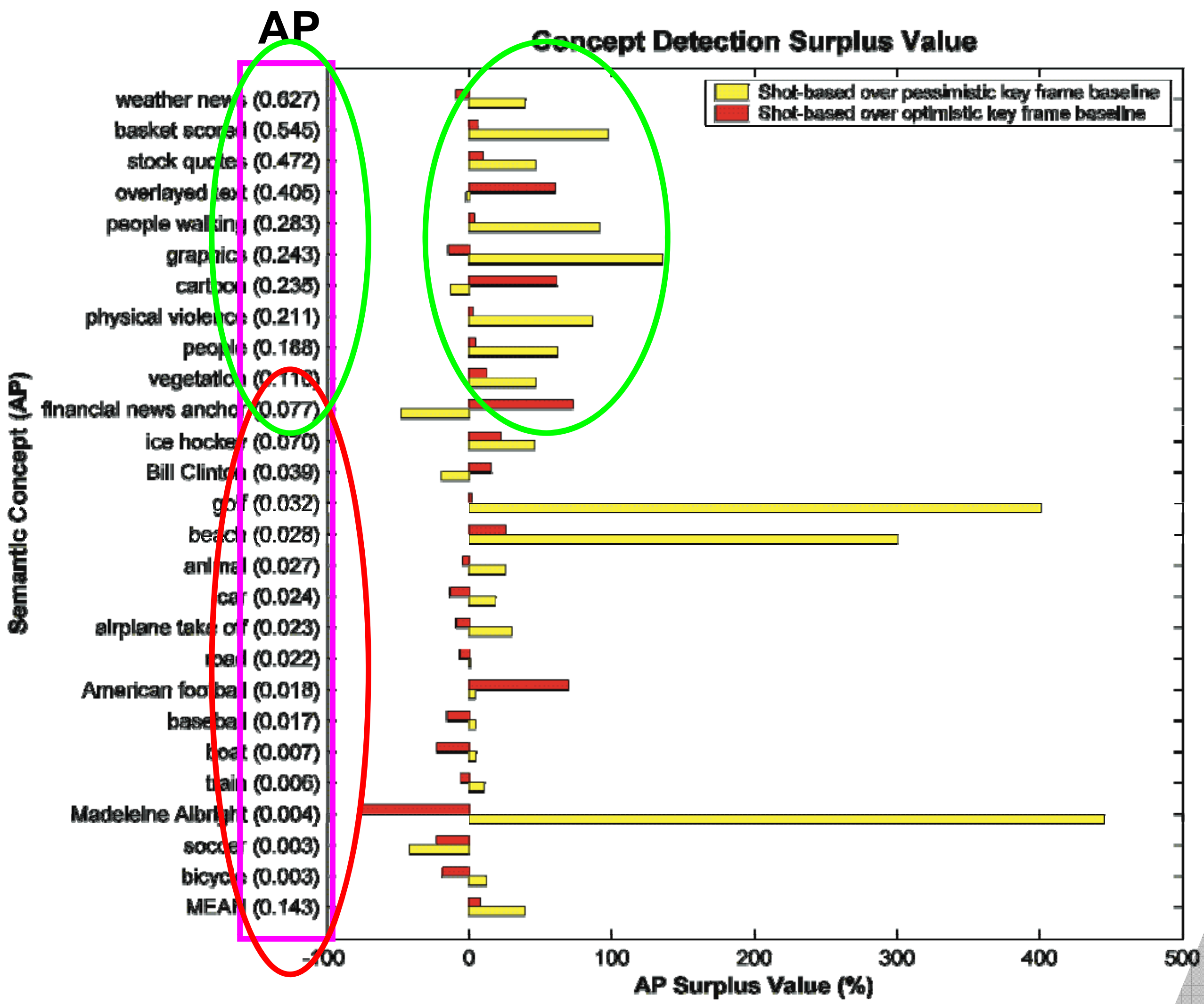
Pessimistic key frame baseline: arg min $P(car \mid frame)$

Optimistic key frame baseline: arg max $P(car \mid frame)$

Concept detection surplus value

TRECVID lessons

- What
- Why
- How
- Why not?
- Conclusion



How to analyze large video archives?

- What
- Why
- How
- Why not?
- Conclusion

- Processing beyond the key frame is expensive
  - ✓ Estimated processing on 1 machine: 250 days
  - ✓ Parallel-Horus on Das-2: < 60 hours
- Parallel-Horus
  - ✓ Efficient parallel execution of sequential software
- Dutch supercomputer Das-2 (at that moment)
  - ✓ 200 1-Ghz dual Pentium III nodes
  - ✓ Located at five Dutch Universities

Authoring Metaphor

- How do all these techniques relate to each other?
- Video is produced by an author
- The author departs from a semantic intention …
- … articulated in a (sub)consciously selected style structuring and emphasizing parts of the content …
- … and communicated in context with the audience by a set of shared notions.

Video analysis best is the inversion of the production.

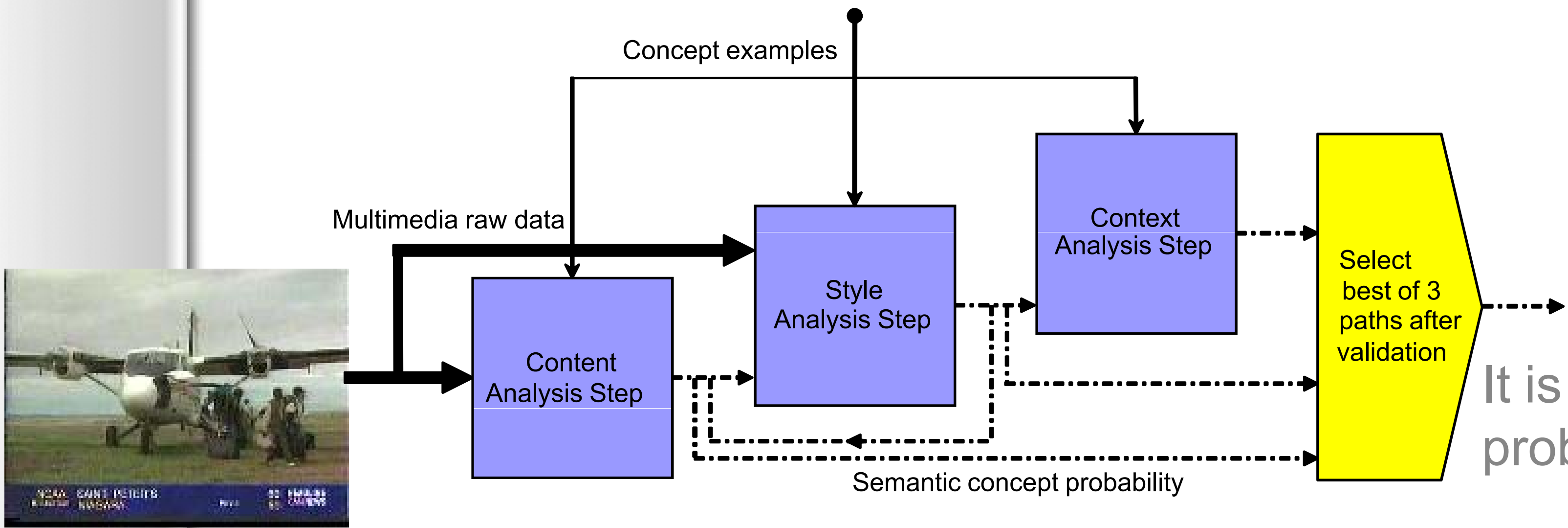
Integrated architecture principled on authoring metaphor

Semantic Pathfinder

following the authoring metaphor

- What
- Why
- How
- Why not?
- Conclusion

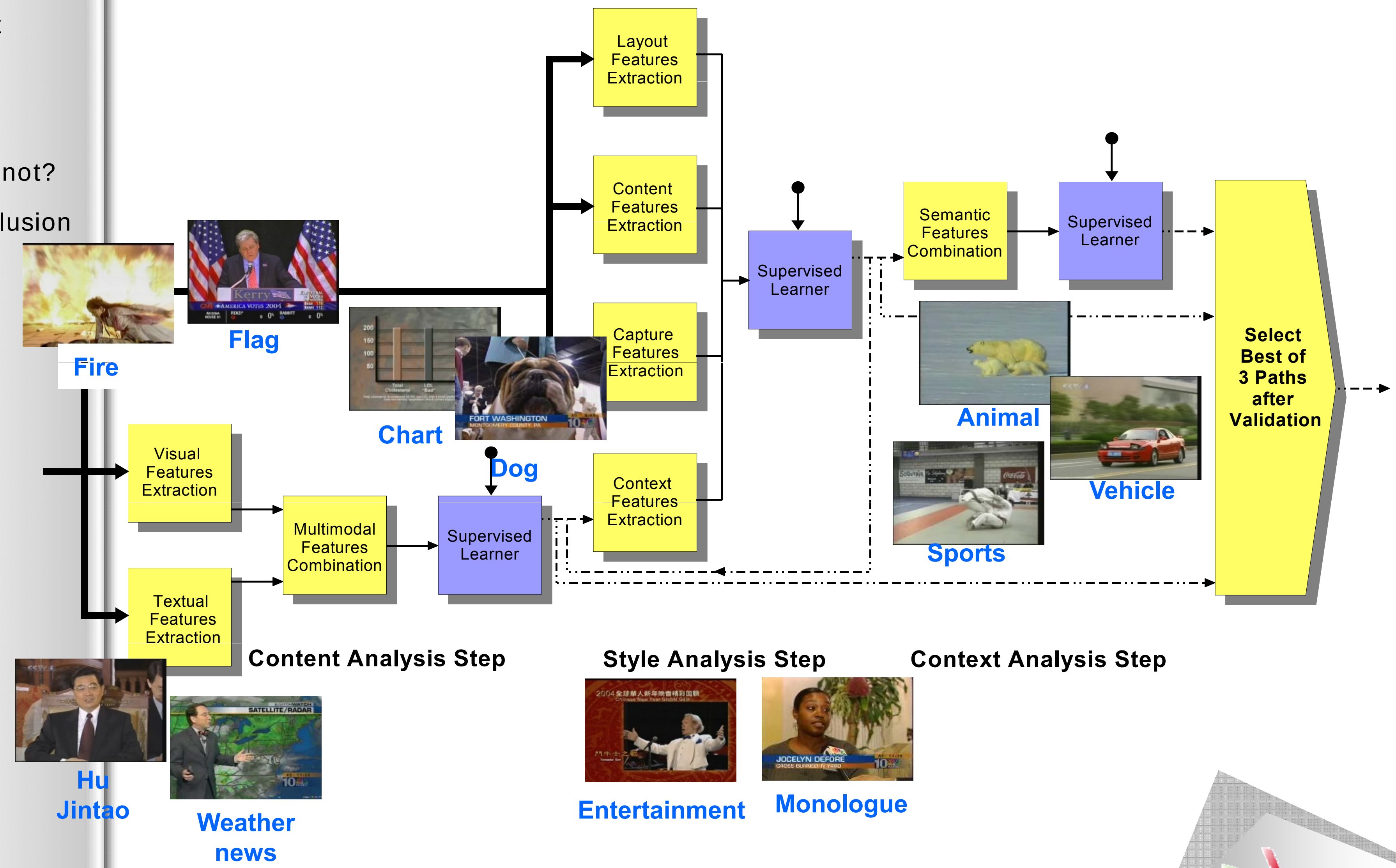




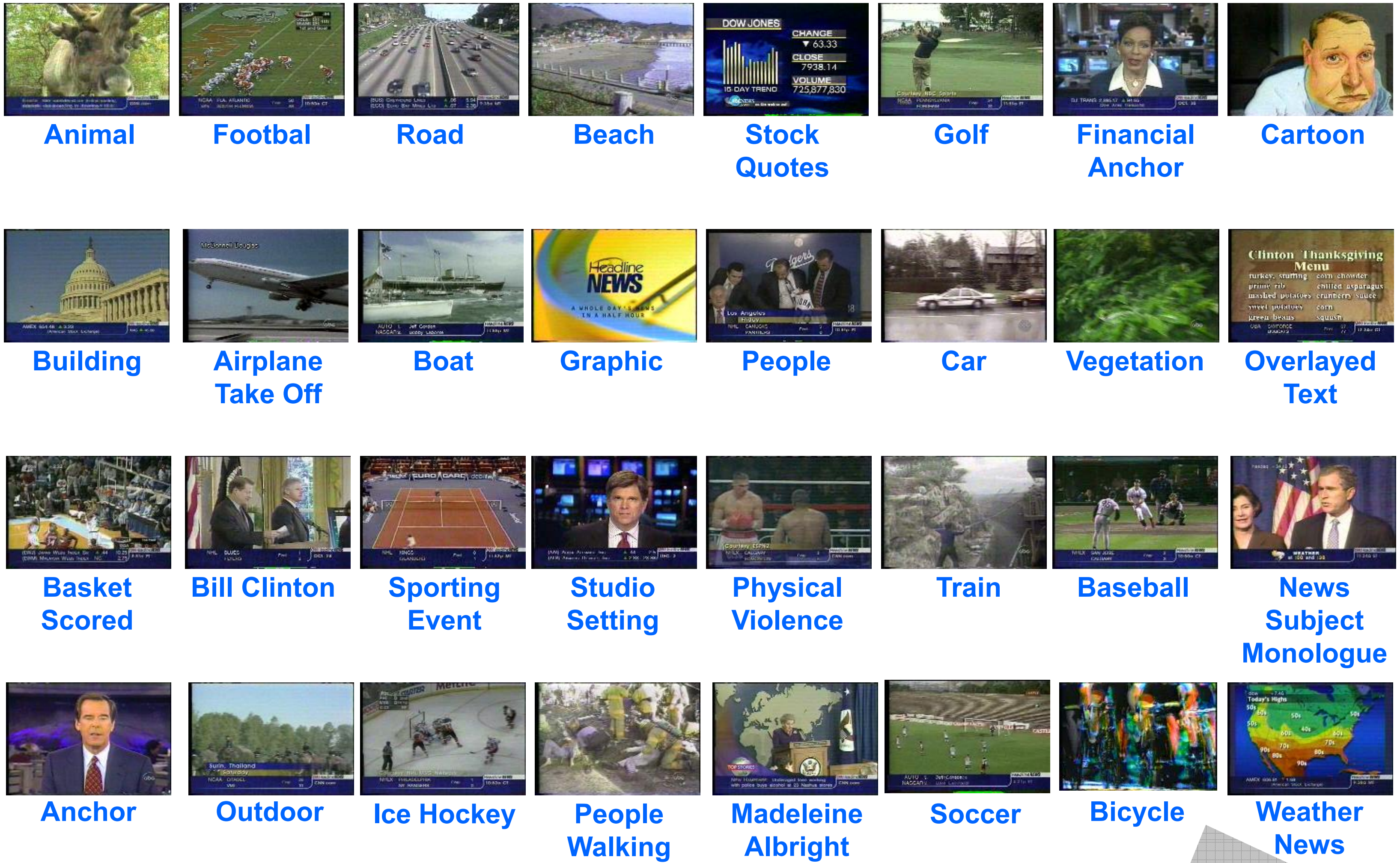
It is an aircraft probability 0.7

Semantic Pathfinder

- What
- Why
- How
- Why not?
- Conclusion



Annotated 32 concept lexicon



Semantic Pathfinder results

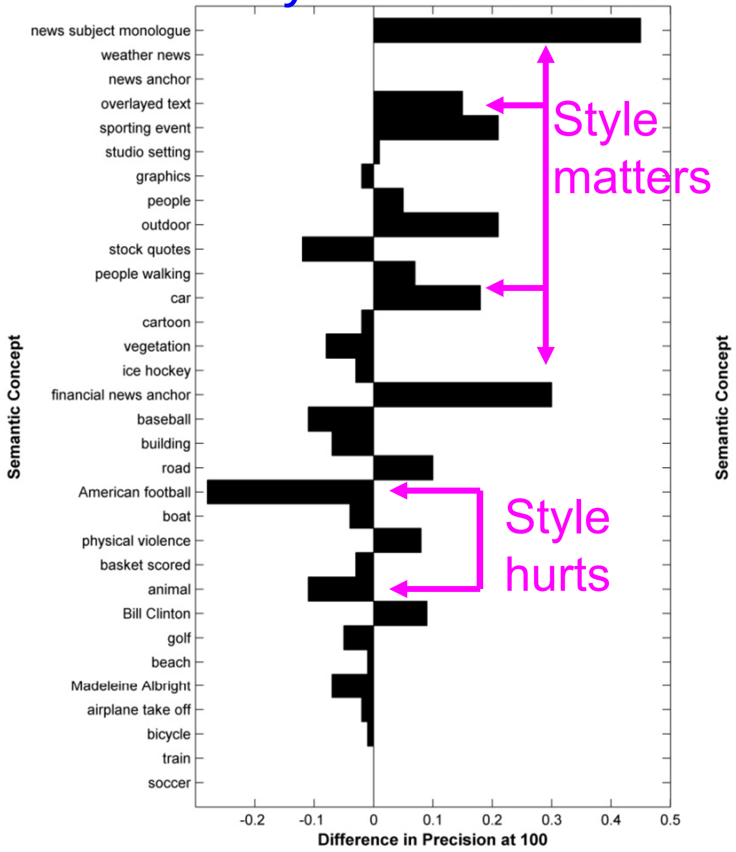
precision@100

	Content	Style	Context	Pathfinder
News subject monologue	0.55	**1.00**	1.00	1.00
Weather news	**1.00**	1.00	1.00	1.00
News anchor	0.98	0.98	**0.99**	0.99
Overlayed text	0.84	**0.99**	0.93	0.99
Sporting event	0.77	**0.98**	0.93	0.98
Studio setting	0.95	0.96	**0.98**	0.98
Graphics	0.92	0.90	**0.91**	0.91
People	0.73	0.78	**0.91**	0.91
Outdoor	0.62	0.83	**0.90**	0.90
Stock quotes	**0.89**	0.77	0.77	0.89
People walking	0.65	0.72	**0.83**	0.83
Car	0.63	0.81	**0.75**	0.75
Cartoon	0.71	0.69	**0.75**	0.75
Vegetation	**0.72**	0.64	0.70	0.72
Ice hockey	**0.71**	0.68	0.60	0.71
Financial news anchor	0.40	**0.70**	0.71	0.70
Baseball	**0.54**	0.43	0.47	0.54
Building	**0.53**	0.46	0.43	0.53
Road	0.42	0.53	**0.51**	0.51
American football	**0.46**	0.18	0.17	0.46
Boat	0.42	0.38	**0.37**	0.37
Physical violence	0.17	0.25	**0.31**	0.31
Basket scored	0.24	0.21	**0.30**	0.30
Animal	0.37	0.26	**0.26**	0.26
Bill Clinton	**0.26**	0.35	0.37	0.26
Golf	**0.24**	0.19	0.06	0.24
Beach	0.13	0.12	**0.12**	0.12
Madeleine Albright	**0.12**	0.05	0.04	0.12
Airplane take off	0.10	0.08	**0.08**	0.08
Bicycle	0.09	**0.08**	0.07	0.08
Train	**0.07**	0.07	0.03	0.07
Soccer	**0.01**	0.01	0.00	0.01
Mean	*0.51*	*0.53*	*0.54*	*0.57*

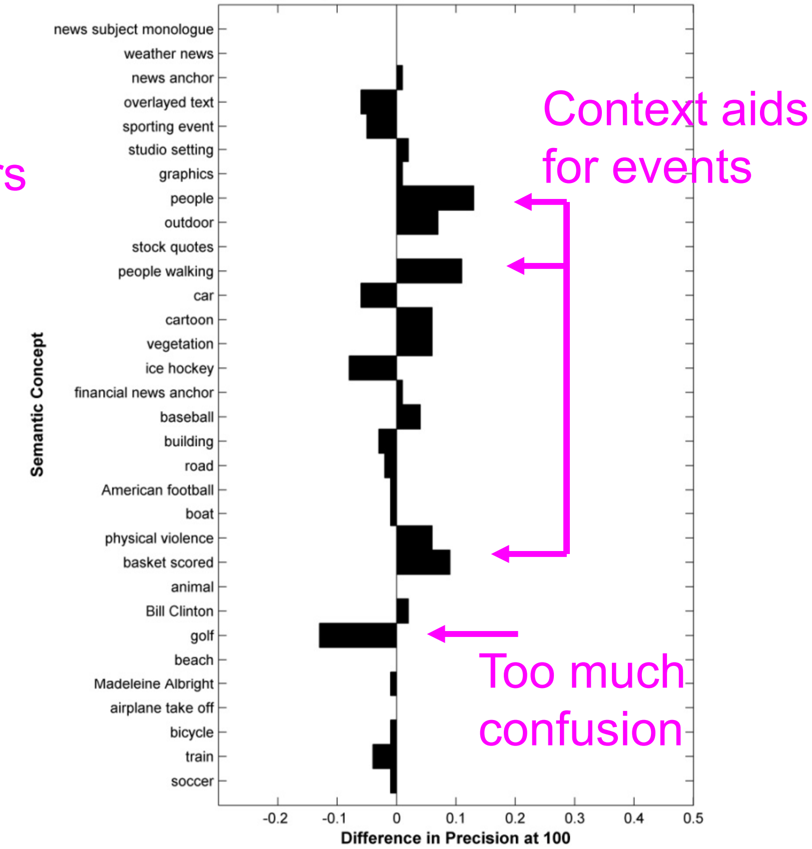
Influence of style & context

precision@100

Style Influence



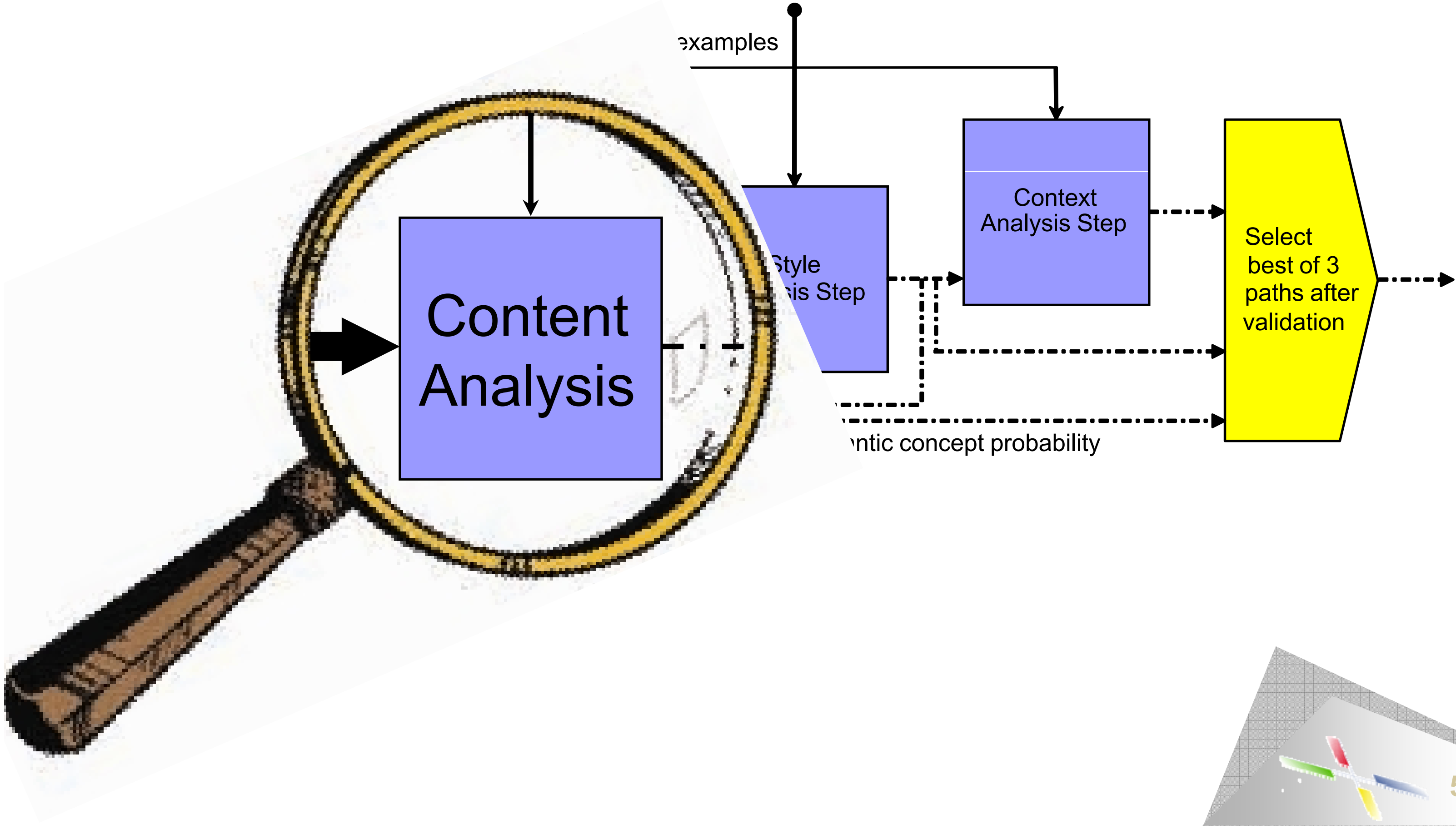
Context Influence



Content analysis pathfinder

TRECVID 2005

- Further refinement of semantic pathfinder
  - ✓ Emphasizing content analysis step
  - ✓ Are some concepts visual, others text, or multimodal?

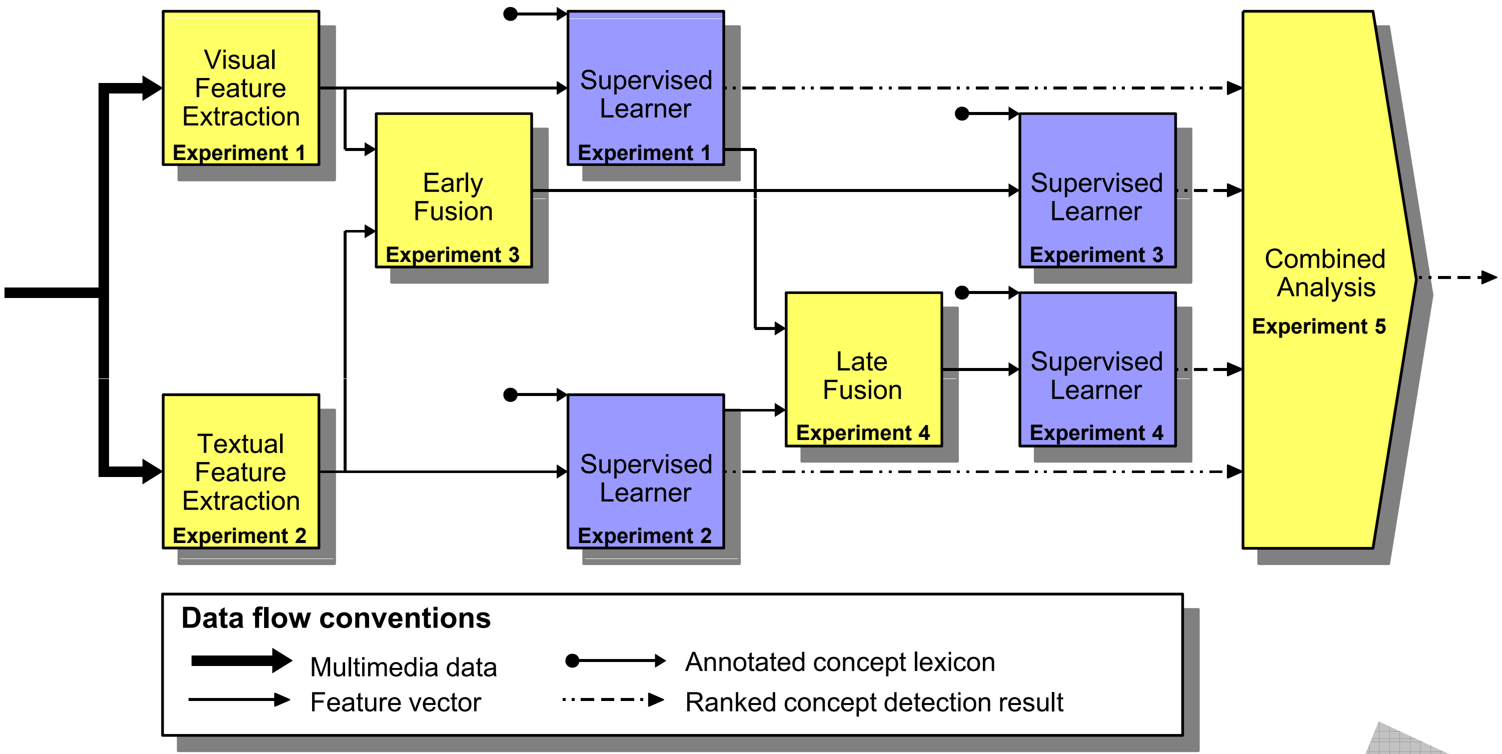


Content analysis pathfinder

TRECVID 2005

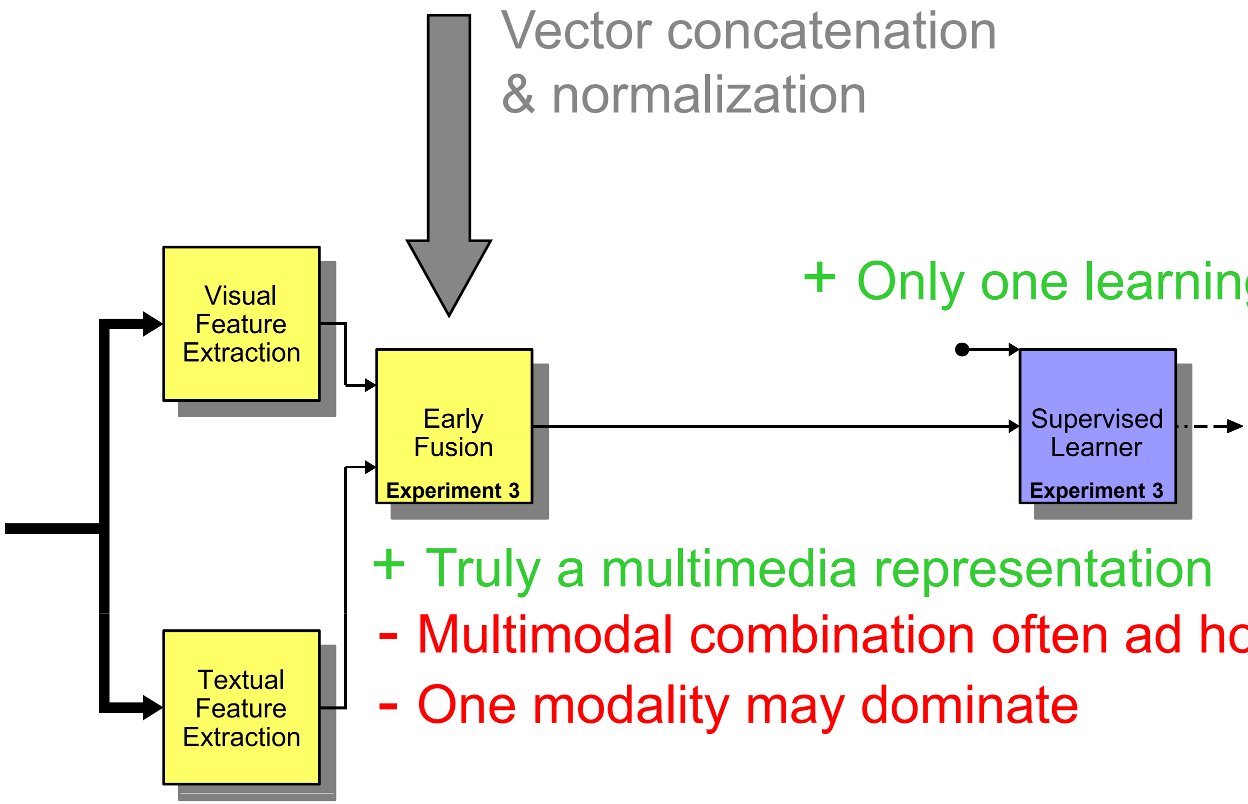
- What
- Why
- How
- Why not?
- Conclusion

➢ Vary unimodal and multimodal combinations



Early fusion

Intermezzo



\+ Only one learning phase

\+ Truly a multimedia representation

\- Multimodal combination often ad hoc

\- One modality may dominate

Late fusion

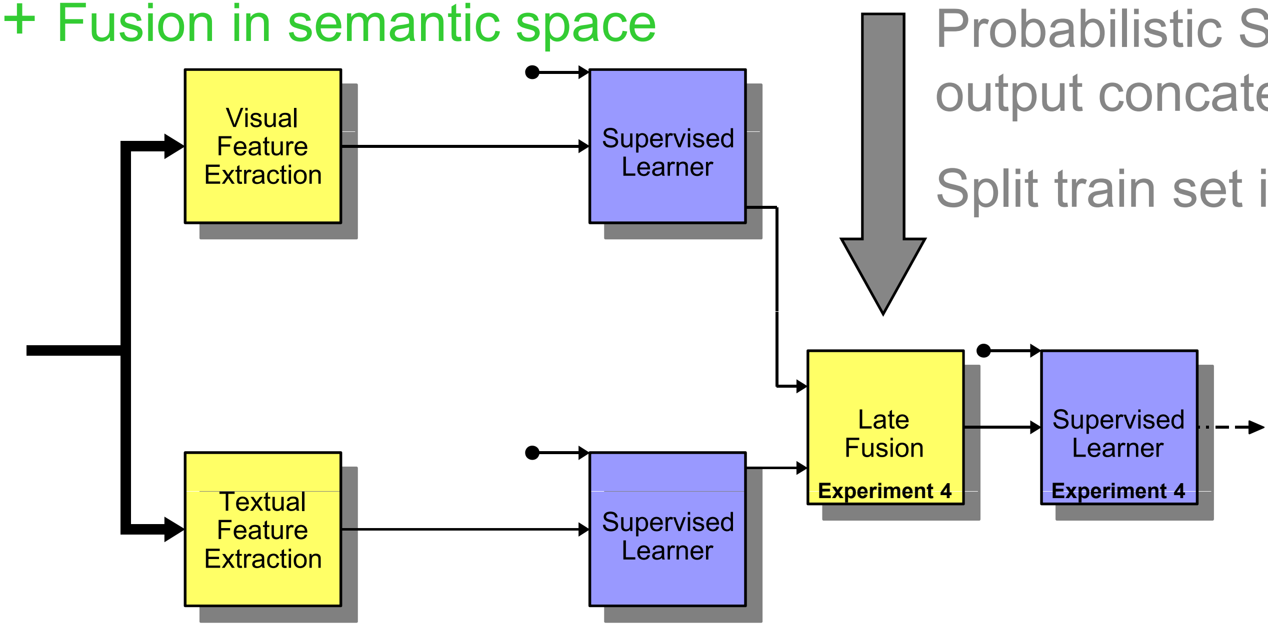
Intermezzo

\+ Focus on modality strength

\+ Fusion in semantic space

Probabilistic SVM output concatenation

Split train set in two



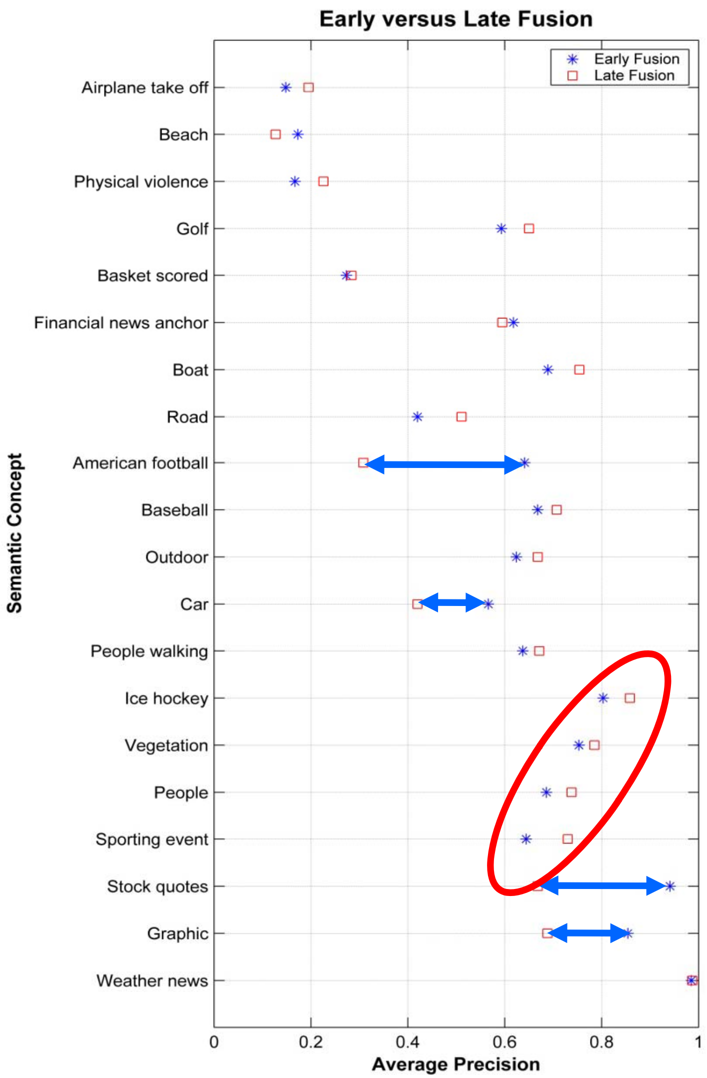
\- Expensive in terms of learning effort

\- Possible loss of feature space correlation

Early vs Late Fusion

- What
- Why
- How
- Why not?
- Conclusion

Intermezzo



□ Late Fusion

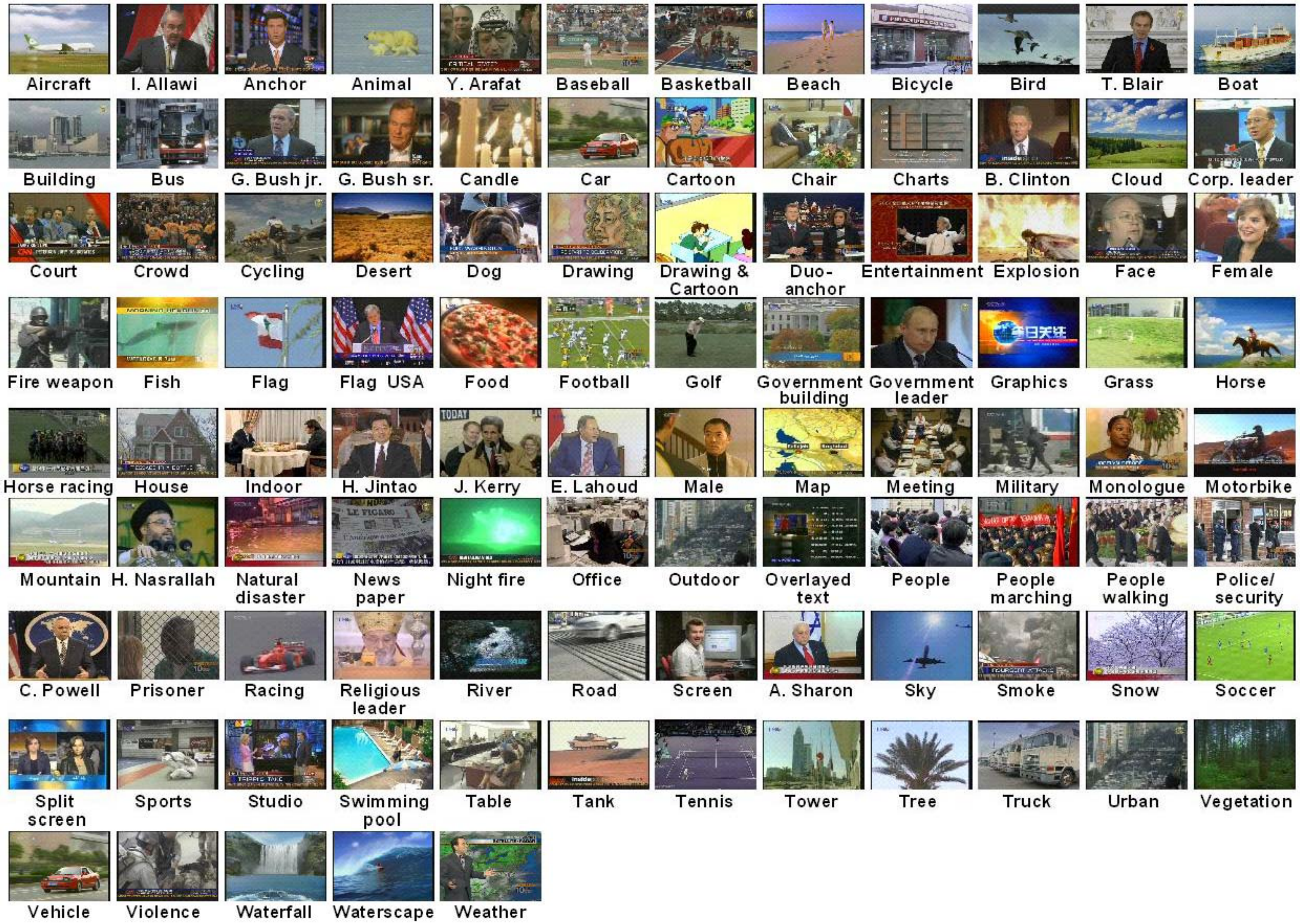
* Early Fusion

- Late Fusion
  - ✓ 14x best performer
  - ✓ AP increase [0 – 0.1]
  - ✓ Extra learning aids performance
- Early Fusion
  - ✓ 6x best performer
  - ✓ AP increase [0 – 0.3]
  - ✓ If better, more significant
- Best fusion strategy
  - ✓ concept-dependent

- What
- Why
- How
- Why not?
- Conclusion

Annotated 101 concept lexicon

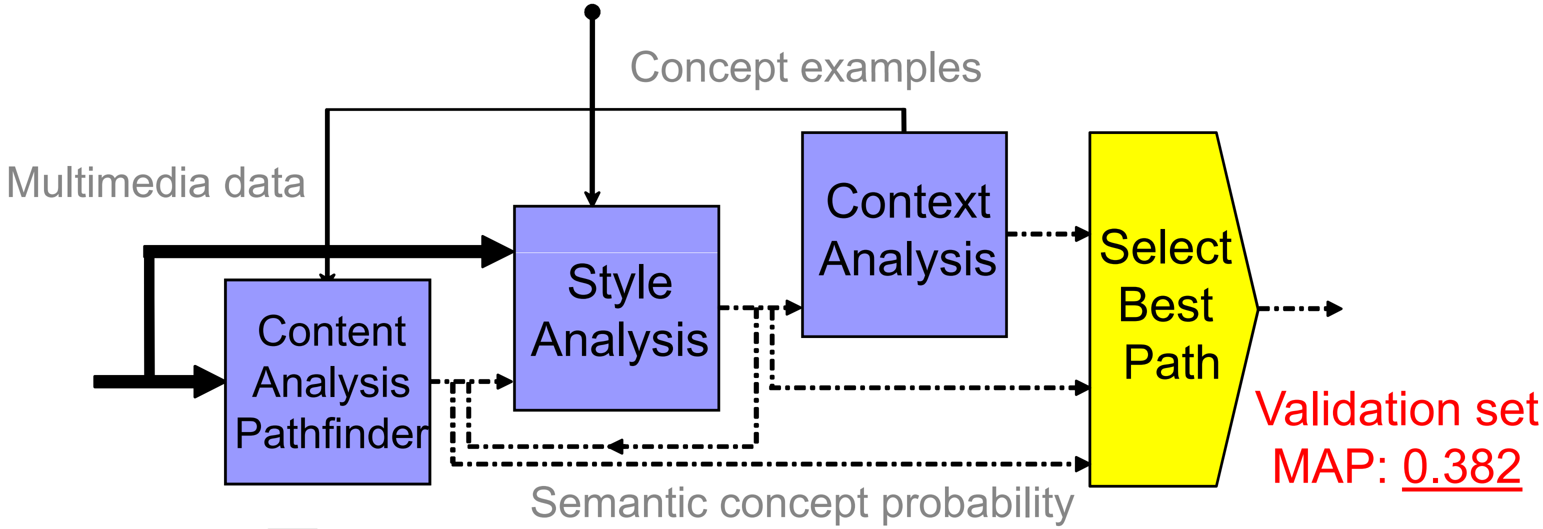
TRECVID 2005



Semantic Pathfinder

- What
- Why
- How
- Why not?
- Conclusion

TRECVID 2005



Validation set
MAP: 0.298

Validation set
MAP: 0.263

Validation set
MAP: 0.352

Animal

Vehicle

Sports

Anchor

Entertainment

Monologue

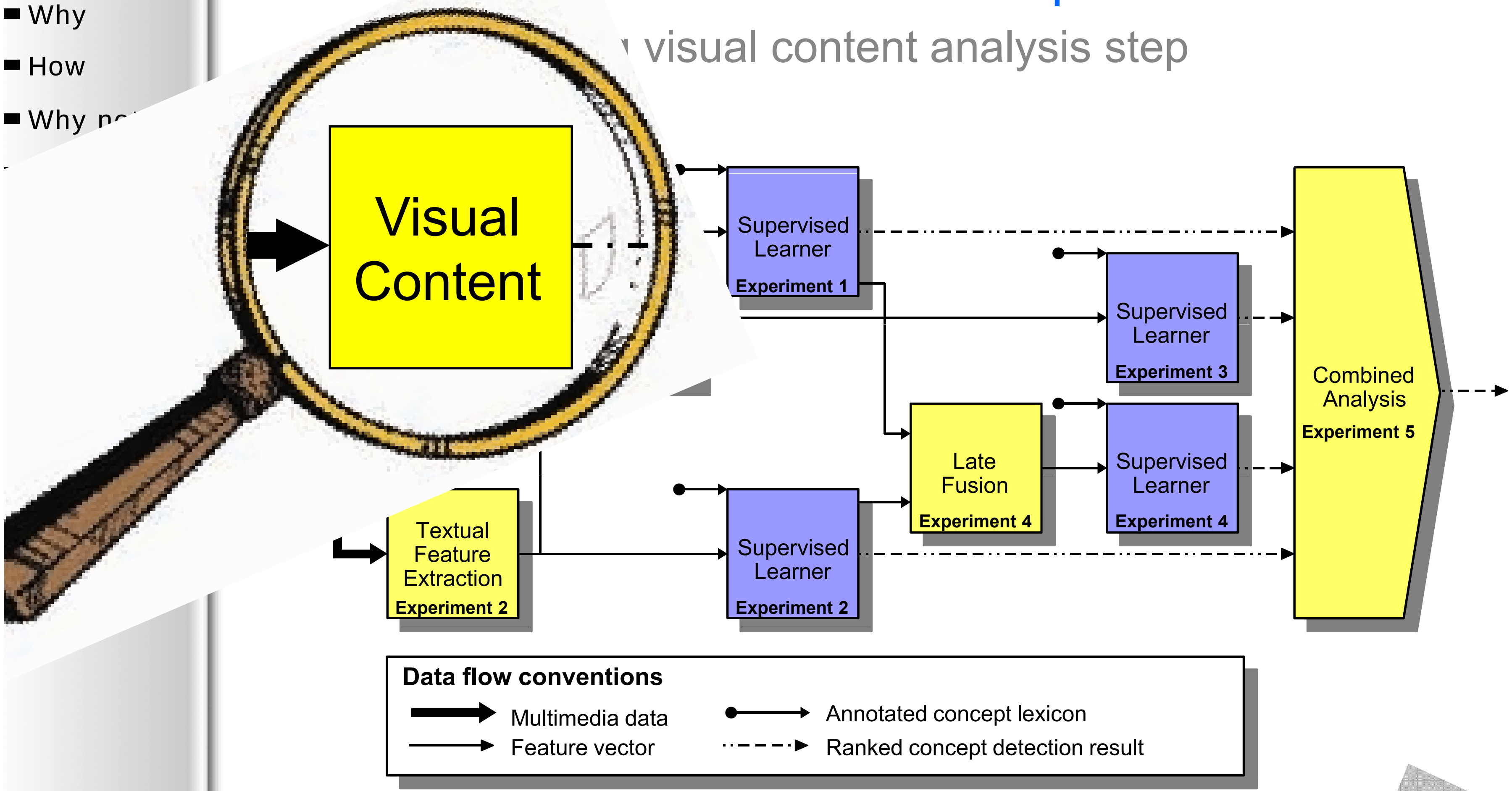
Visual content analysis pathfinder

TRECVID 2006

- What
- Why
- How
- Why n

➢ Furthe ement of semantic pathfinder

visual content analysis step



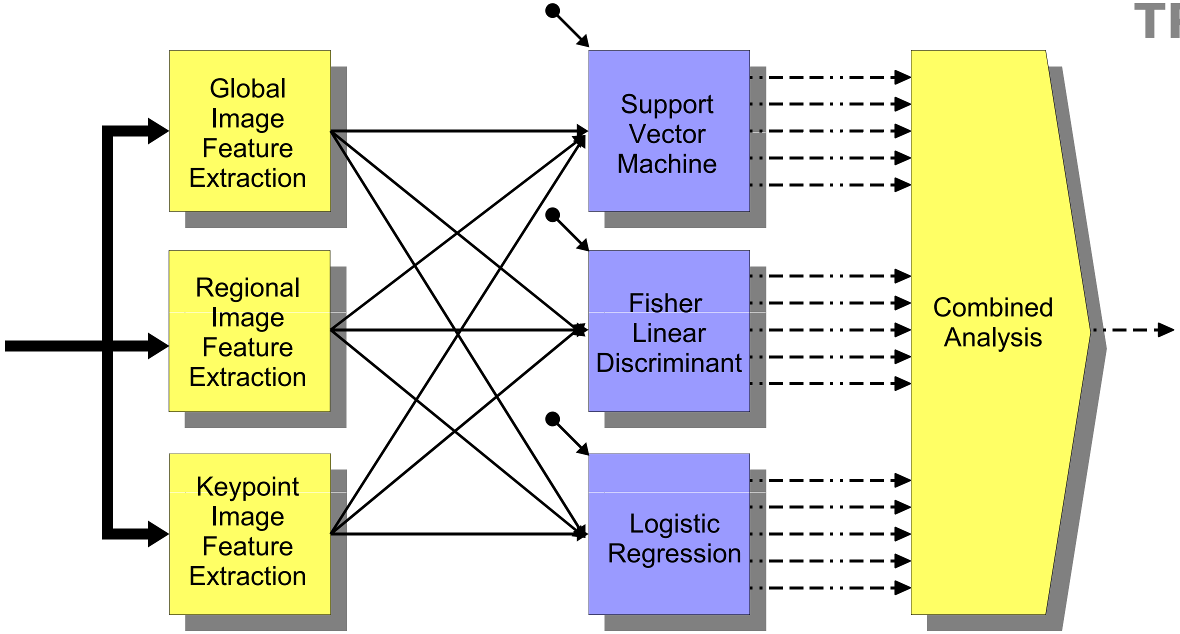
Annotated 491 concept lexicon

TRECVID 2006

- What
- Why
- How
- Why not?
- Conclusion

Visual content analysis pathfinder

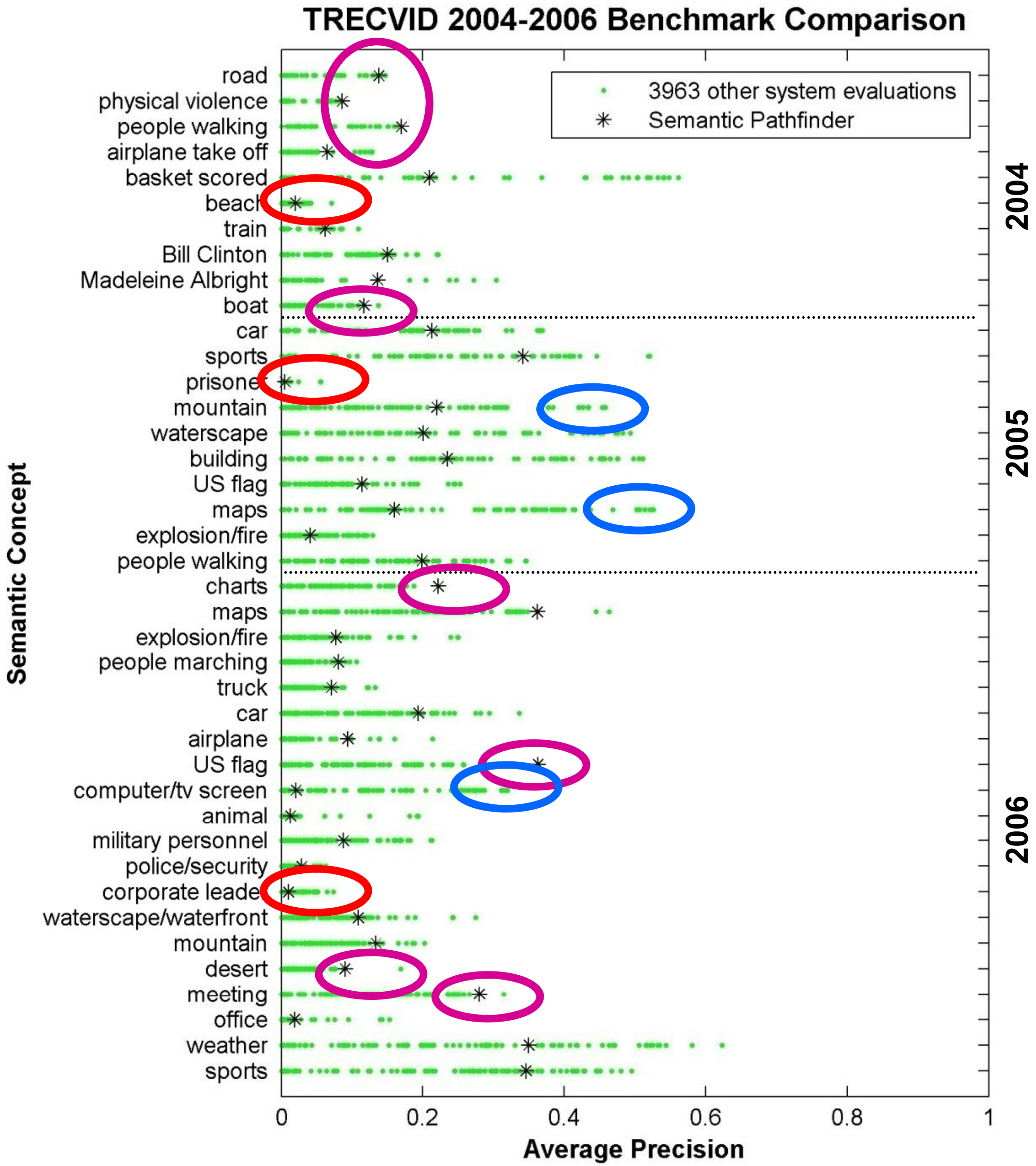
TRECVID 2006



- For visual-only semantic pathfinder we learned that
  - ✓ A combination of various (invariant) visual-only techniques pays off
  - ✓ Regional image features seem most effective
  - ✓ Keypoint methods unstable for images with few interest points
  - ✓ High-dimensional feature vectors can be handled effectively by relatively simple classifiers like Fishers linear discriminant
  - ✓ Fusion using geometric mean is cheap and effective

Semantic Pathfinder @ TRECVID

- What
- Why
- How
- Why not?
- Conclusion



The Good

The Bad

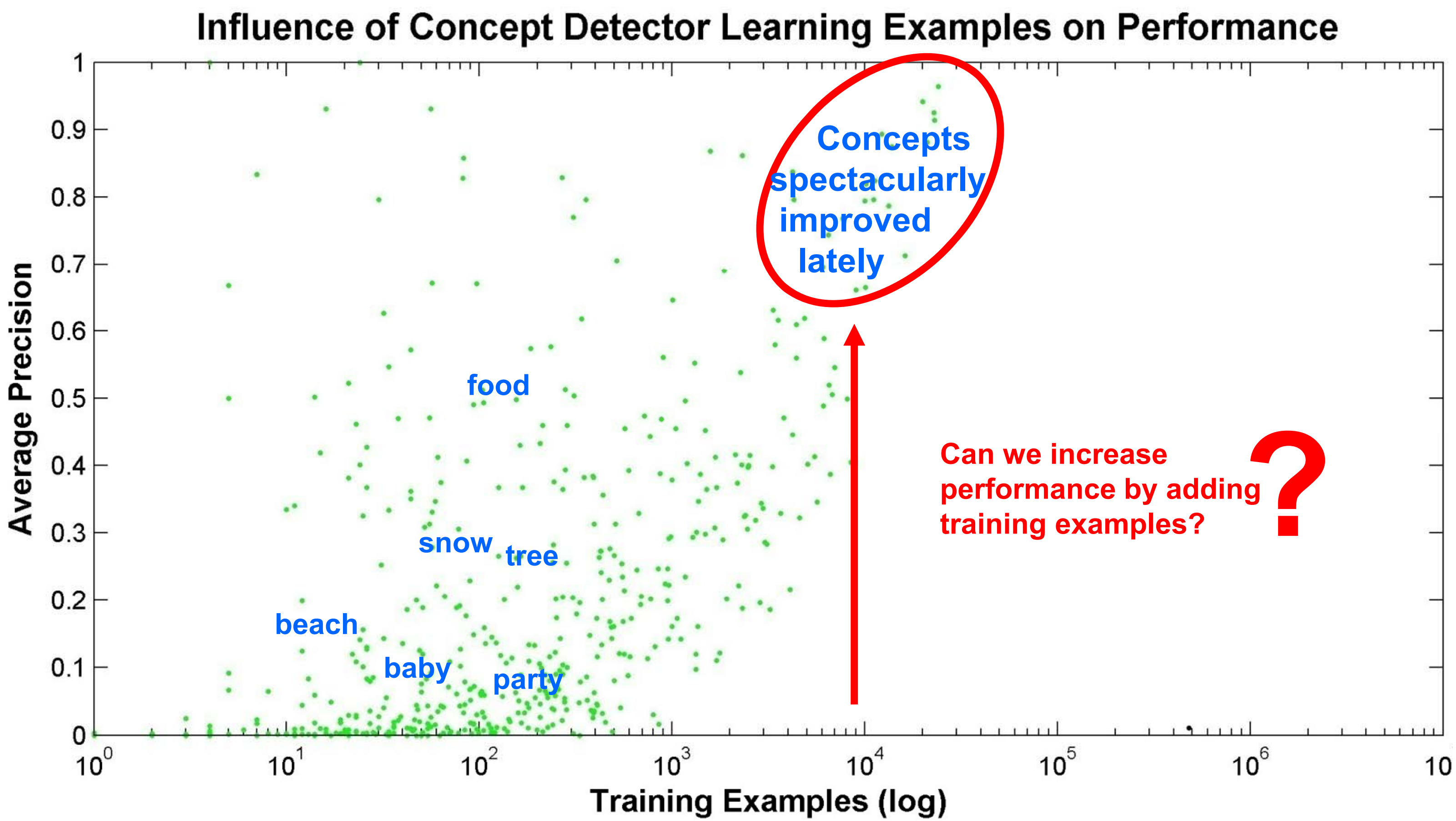
ill-defined / few examples

The Ugly

exploit TV repetition

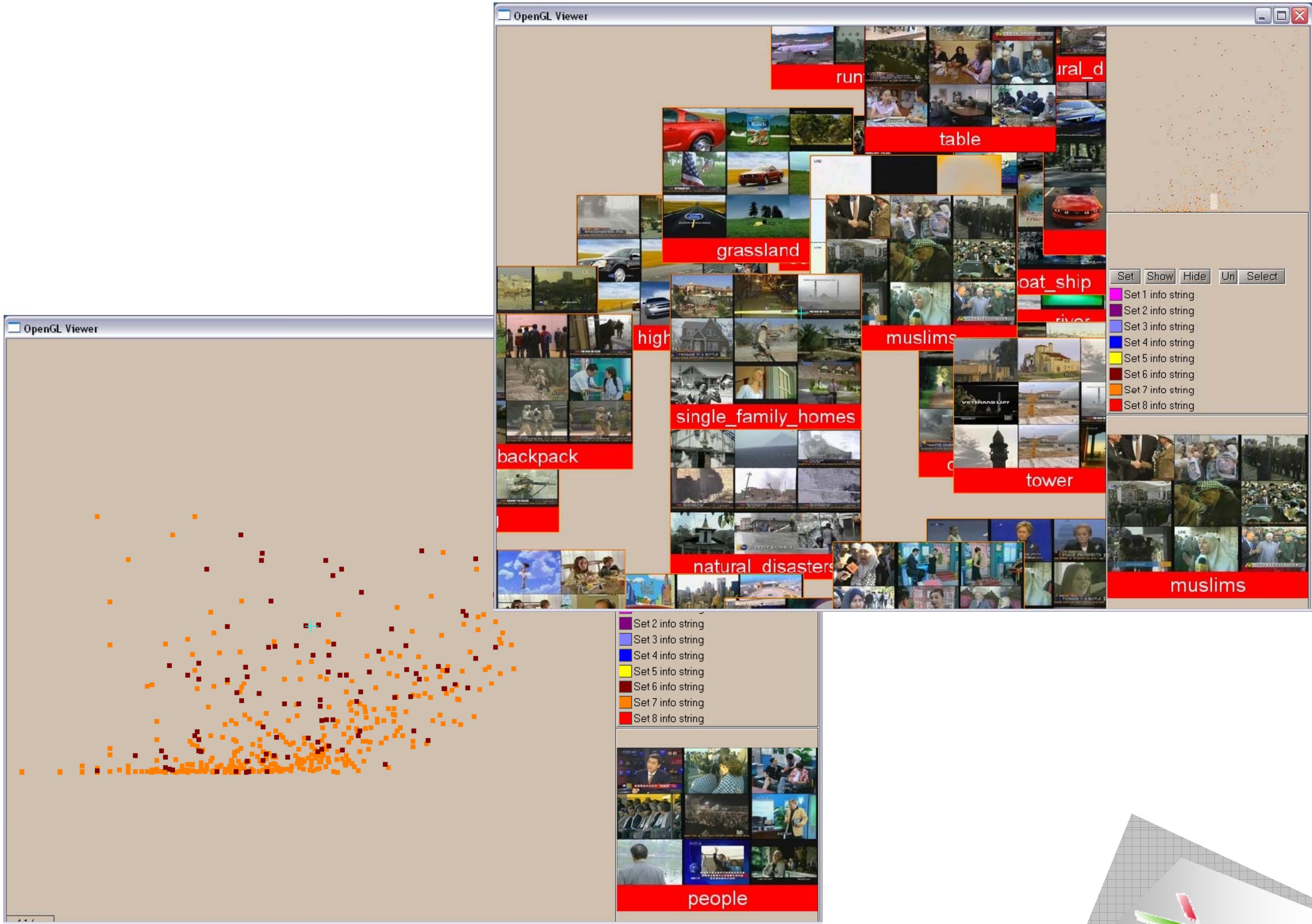
491 detectors, a closer look

- What
- Why
- How
- Why not?
- Conclusion



- What
- Why
- How
- Why not?
- Conclusion

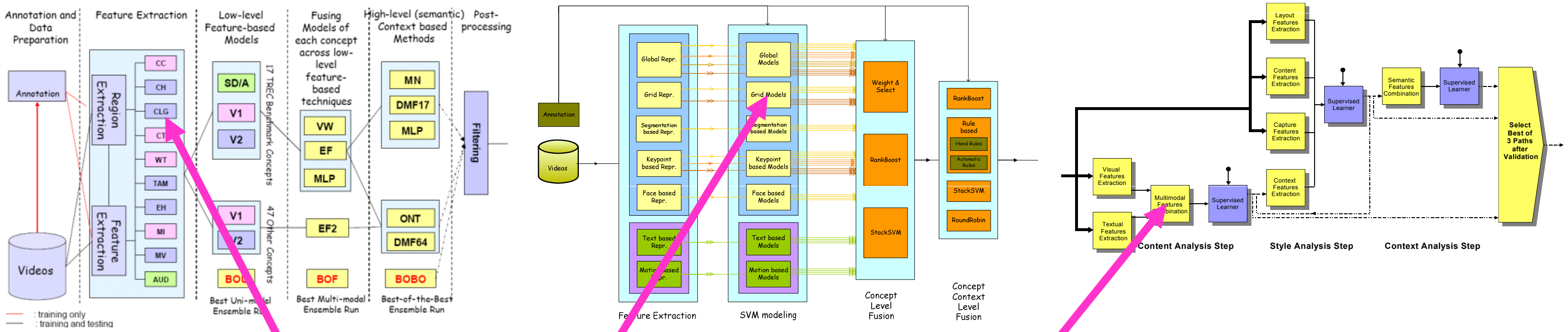
Demo time!



TRECVID Criticism

- What
- Why
- How
- Why not?
- Conclusion

- 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?

- What
- Why
- How
- Why not?
- Conclusion

MediaMill Challenge

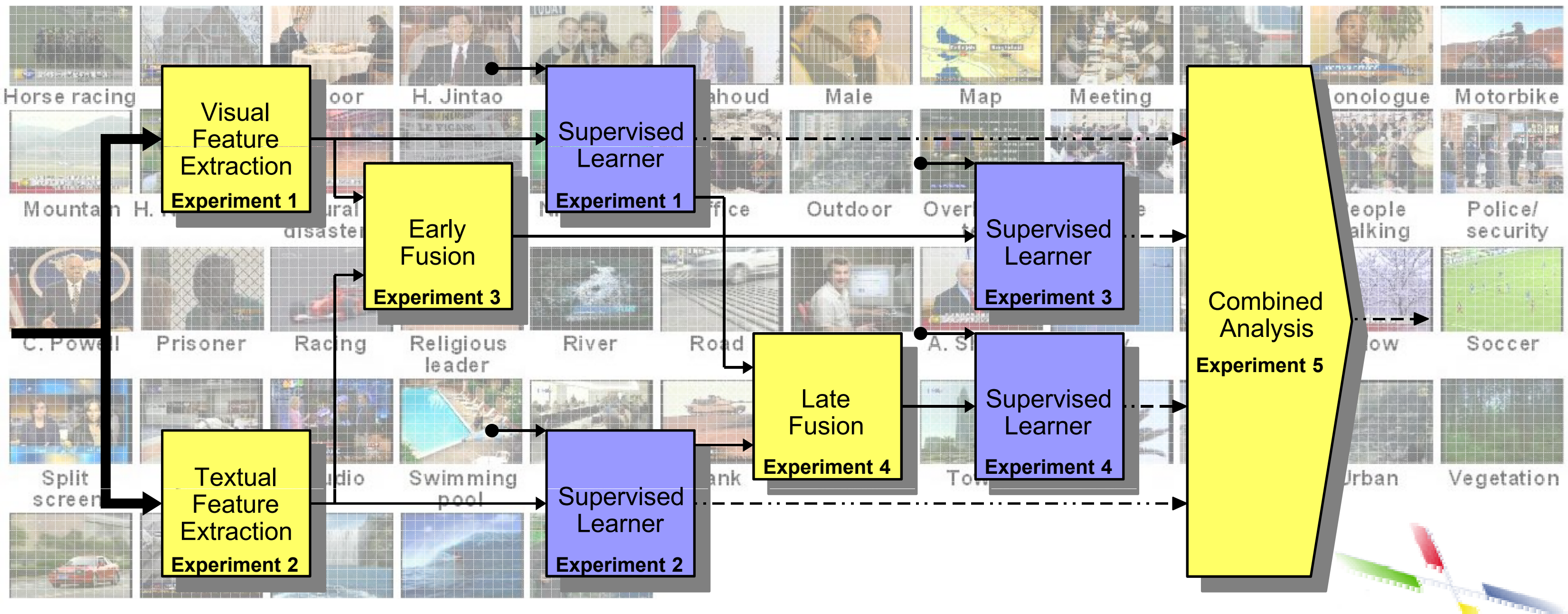
➢ **The Challenge provides**

- ✓ 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/

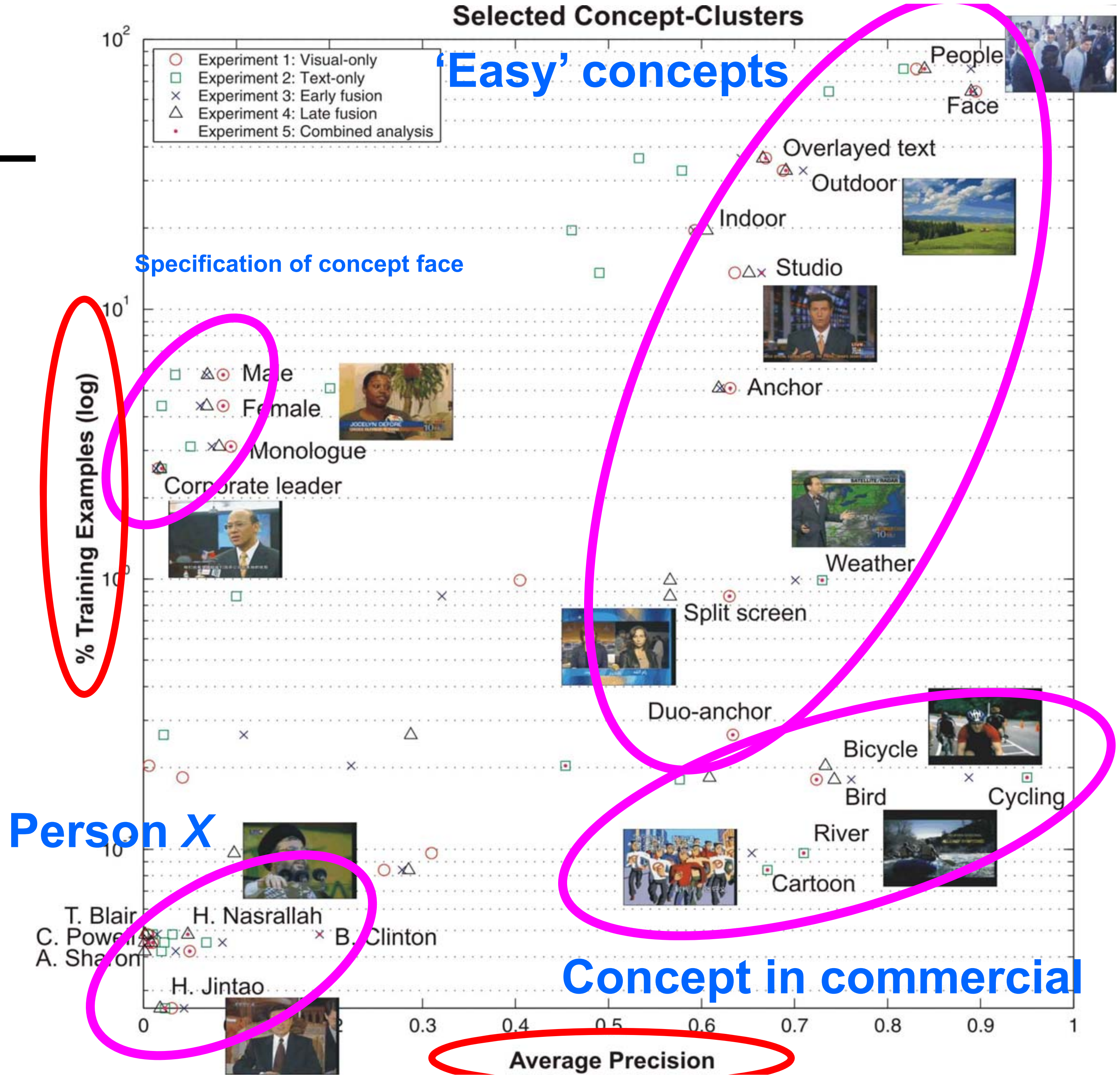
MediaMill Challenge

- What
- Why
- How
- Why not?
- Conclusion

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

- What
- Why
- How
- Why not?
- Conclusion

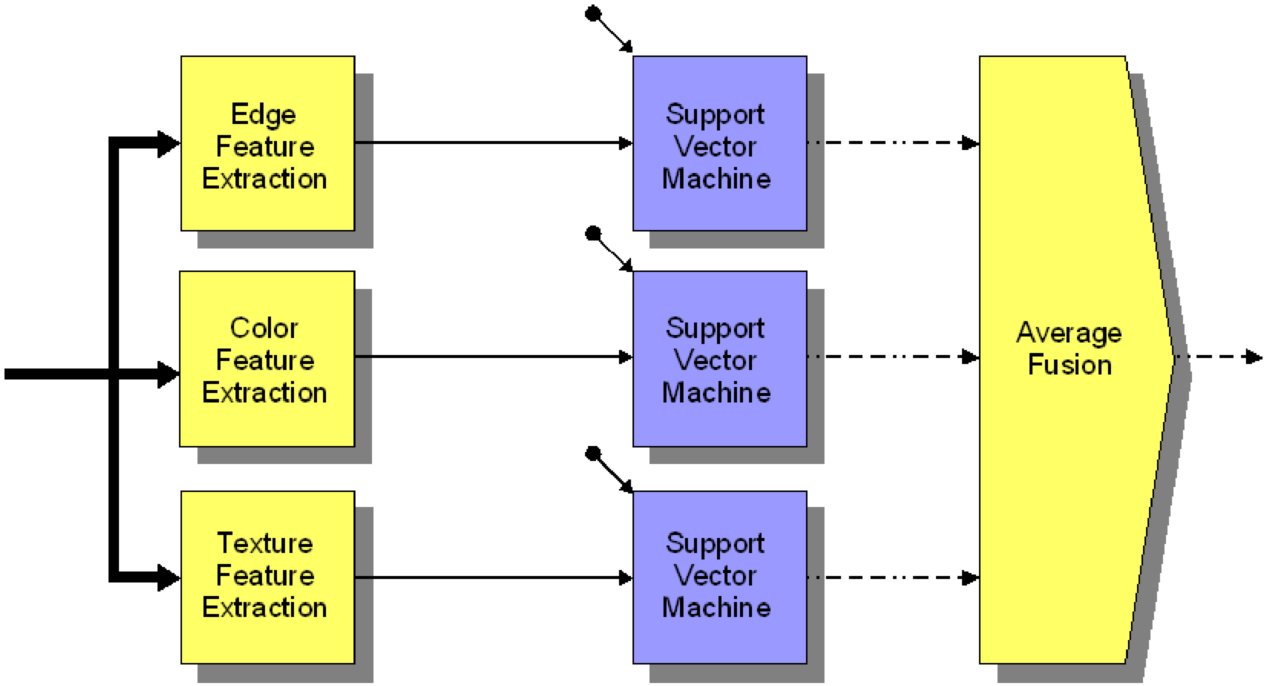


Columbia374

- What
- Why
- How
- Why not?
- Conclusion

➢ Baseline for 374 concept detectors

✓ Focus is on visual analysis experiments

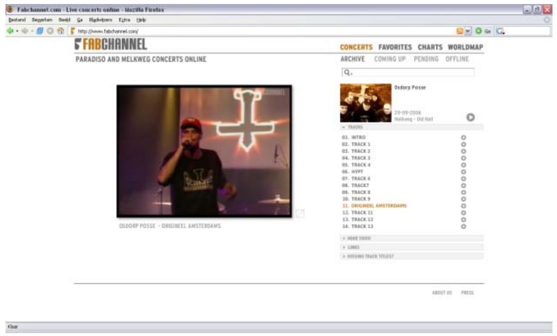


Online available:
http://www.ee.columbia.edu/ln/dvmm/columbia374/

- What
- Why
- How
- Why not?
- Conclusion

Case study
Fabchannel.com

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



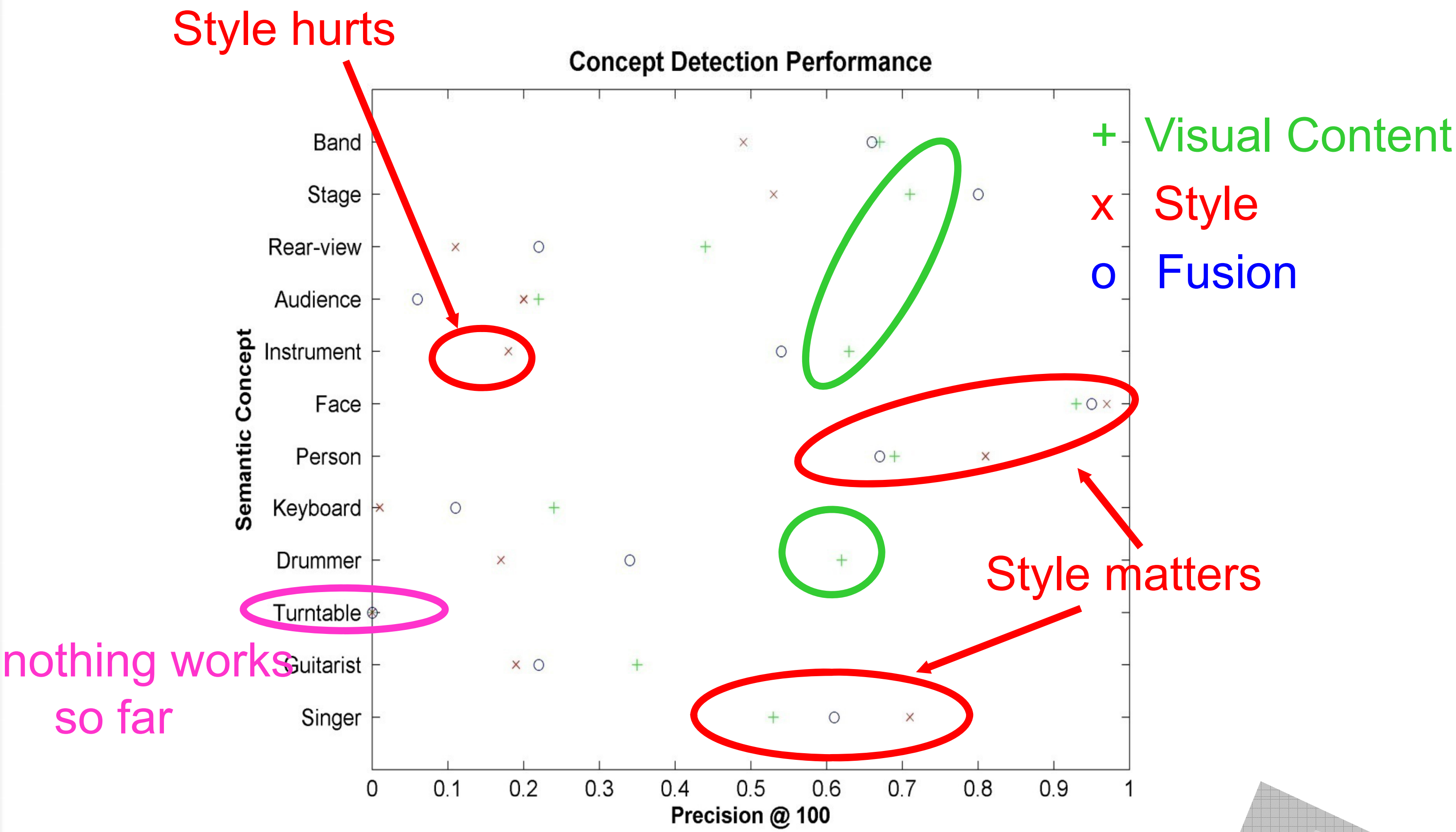
- Fabchannel request
  - ✓ What can you do with 45 hours of live concerts?
- Answer:
  - ✓ Let’s try the semantic pathfinder to detect concert concepts

Case study

Fabchannel.com

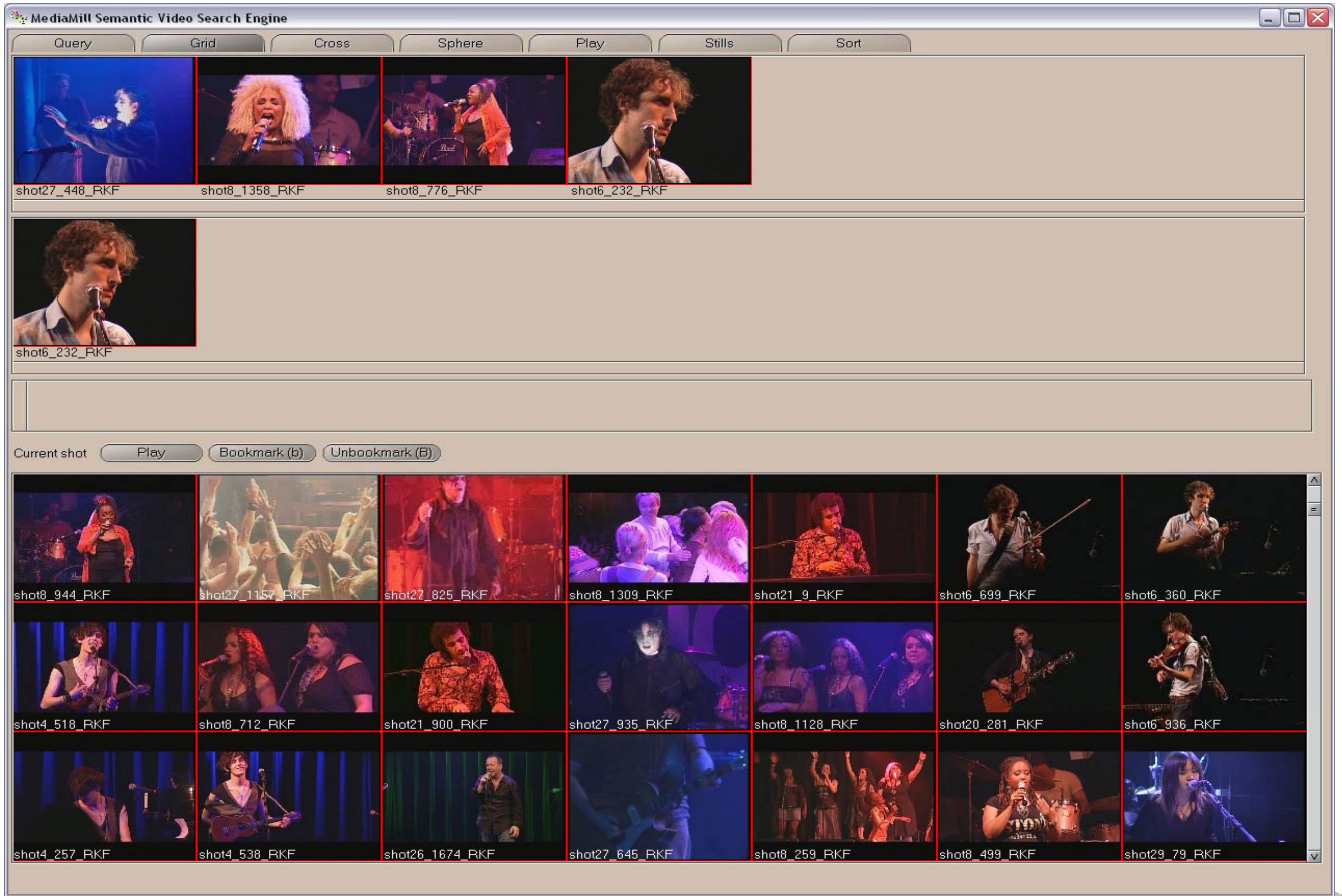
FABCHANNEL LIVE CONCERT VIDEOS

- What
- Why
- How
- Why not?
- Conclusion



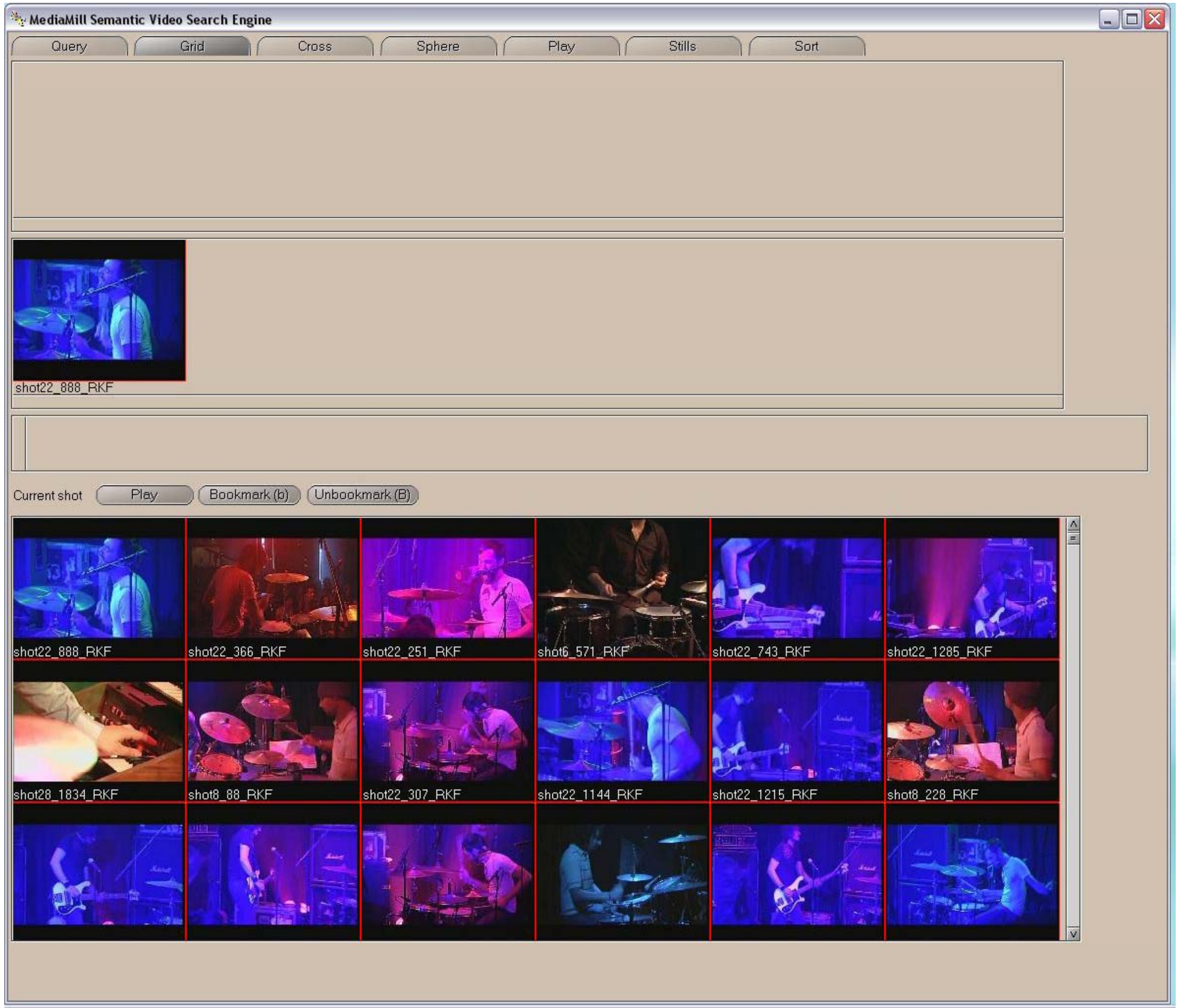
FABCHANNEL
LIVE CONCERT VIDEOS

Results for singer



Results for drummer

- What
- Why
- How
- Why not?
- Conclusion



- What
- Why
- How
- Why not?
- Conclusion

Conclusions

- Semantic pathfinder = generic video indexing
  - ✓ Confirms the authoring metaphor
  - ✓ Currently detects up to ~~101~~ 500 concepts in news video
  - ✓ Generalizes outside news domain
- Technique taxonomy for concept detectors
  - ✓ No superior method for all concepts exists,
  - ✓ Best to learn optimal approach per concept
  - ✓ Some concepts are content, others are style, or context
  - ✓ For content a separation between analysis steps exists also
- State-of-the-art TRECVID performance
  - ✓ Without the need to implement specialized detectors
- Future work
  - ✓ Refinement of pathfinder into people, objects, and setting
  - ✓ Handle sparse learning problem
  - ✓ More feature extraction and classifier schemes?
  - ✓ More annotated data needed!

Concept-based Video Retrieval

Cees Snoek

with contributions by:
many

Intelligent Systems Lab Amsterdam,
University of Amsterdam, The Netherlands

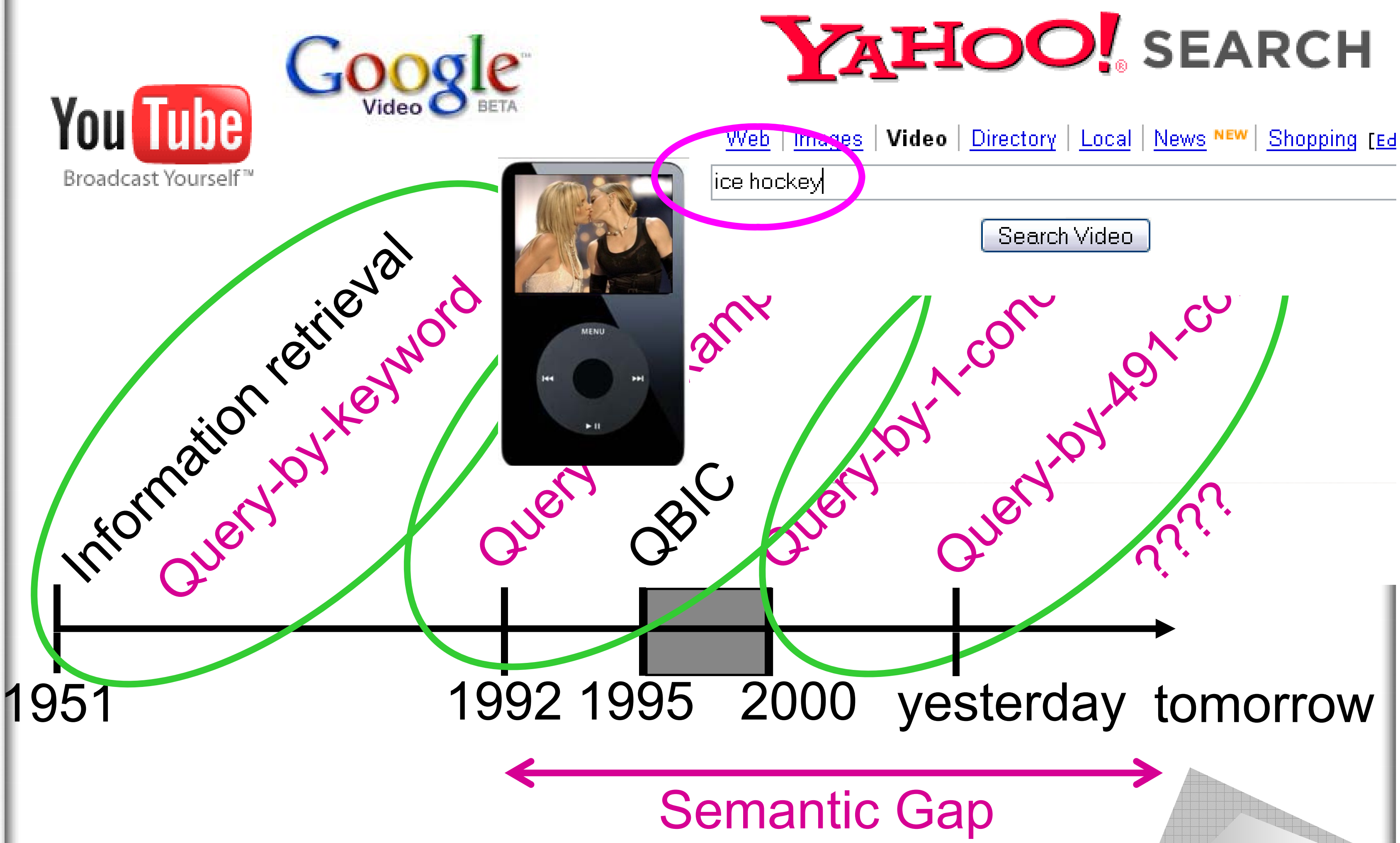
MultimediaN

Why bother?

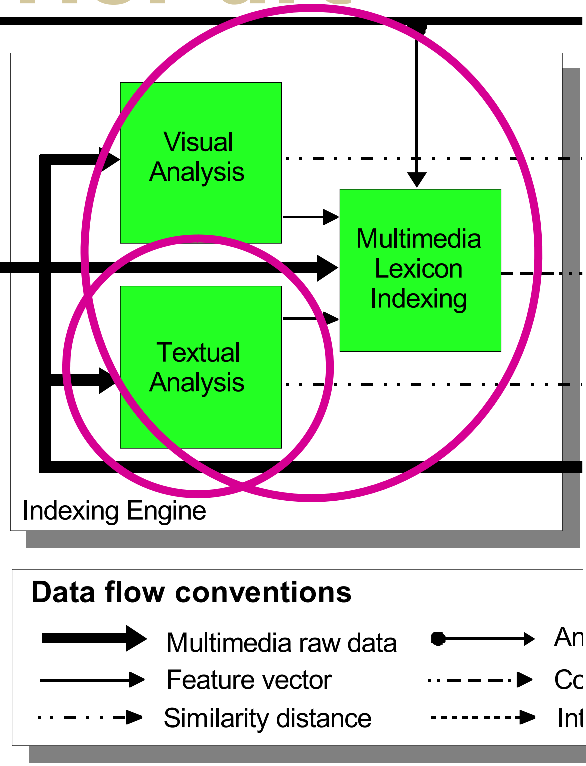
- What
- Why
- How
- Why not?
- Conclusion

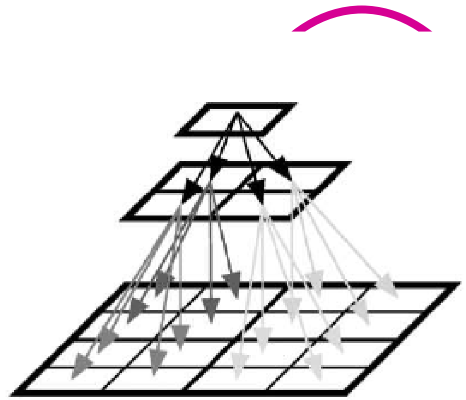
Introduction

- What
- Why
- How
- Why not?
- Conclusion



Prior art





	Keyword	Example	Concept	Lexicon	Display	Evaluation
Adcock	Yes	-	-	0	Story board	Benchmark
Taskiran	-	Yes	Yes	1	Pyramid	Specific
Fan	-	Yes	Yes	5	Hierarchical	Specific
Christel	Yes	Yes	Yes	10	Story board	Benchmark
Smith	Yes	Yes	Yes	17	Grid	Benchmark

MediaMagic

➢ Focus on the story level

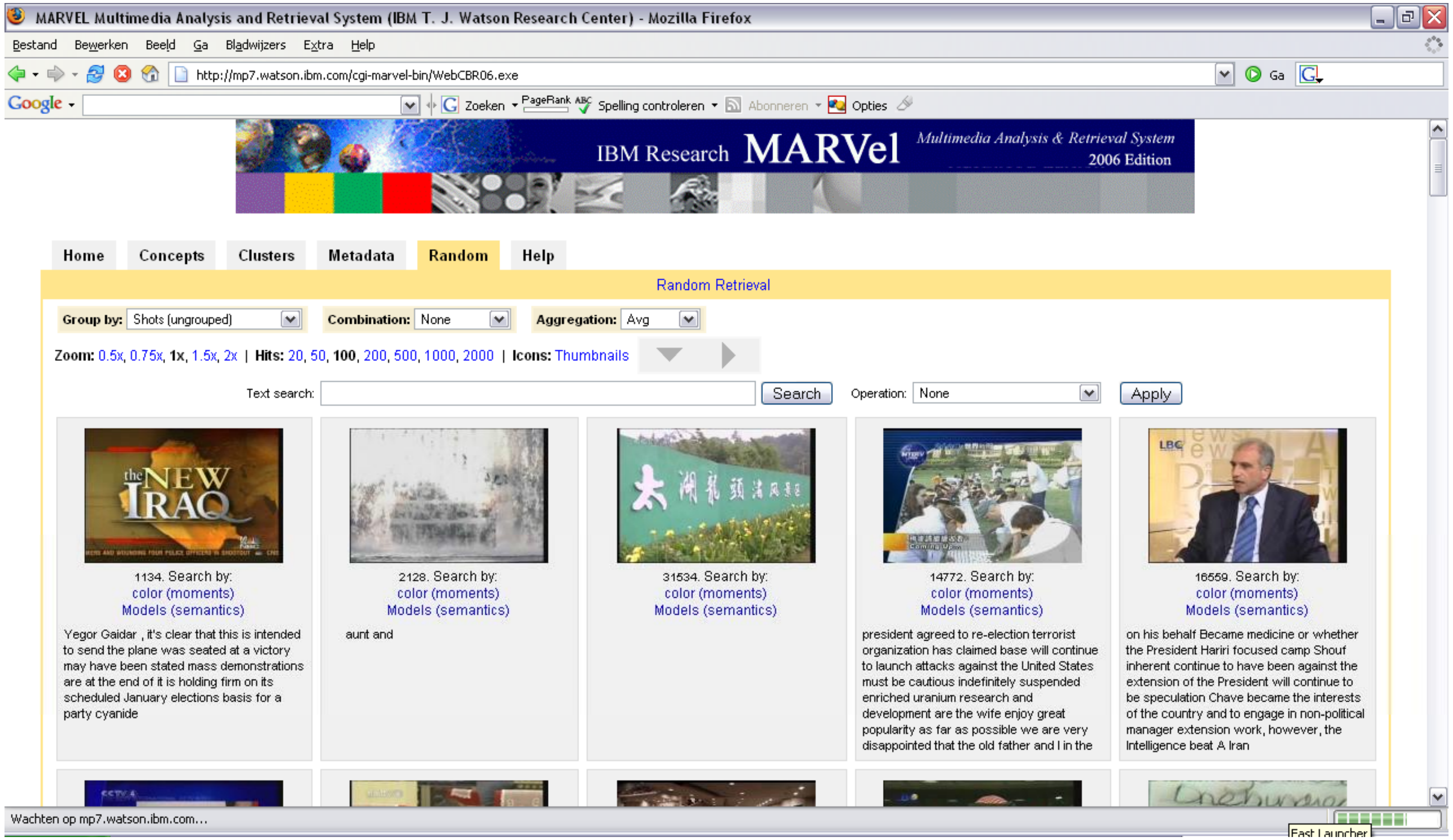
‘Classic’ Informedia system

- First multimodal video search engine

IBM MARVel

- What
- Why
- How
- Why not?
- Conclusion

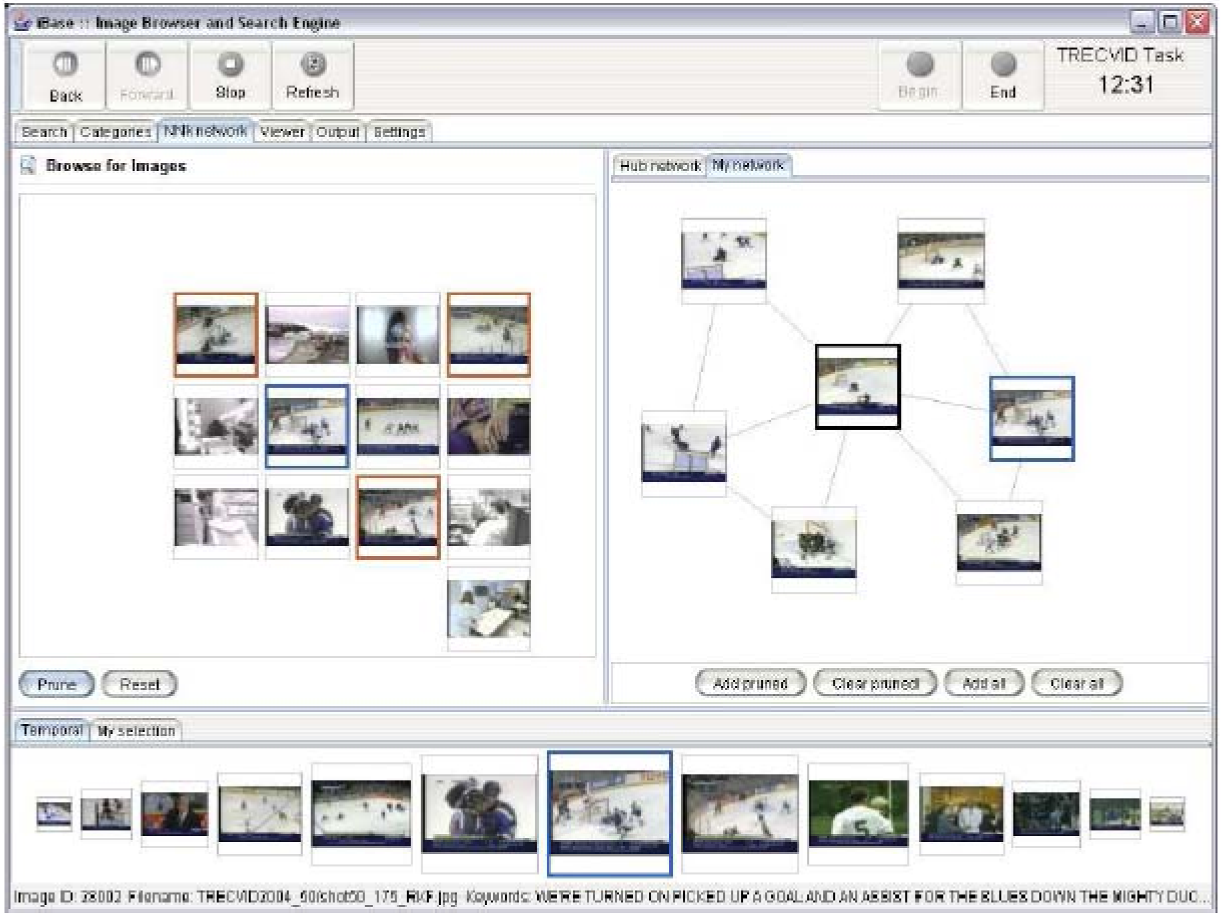
➤ A web based system



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

NN^k Browser

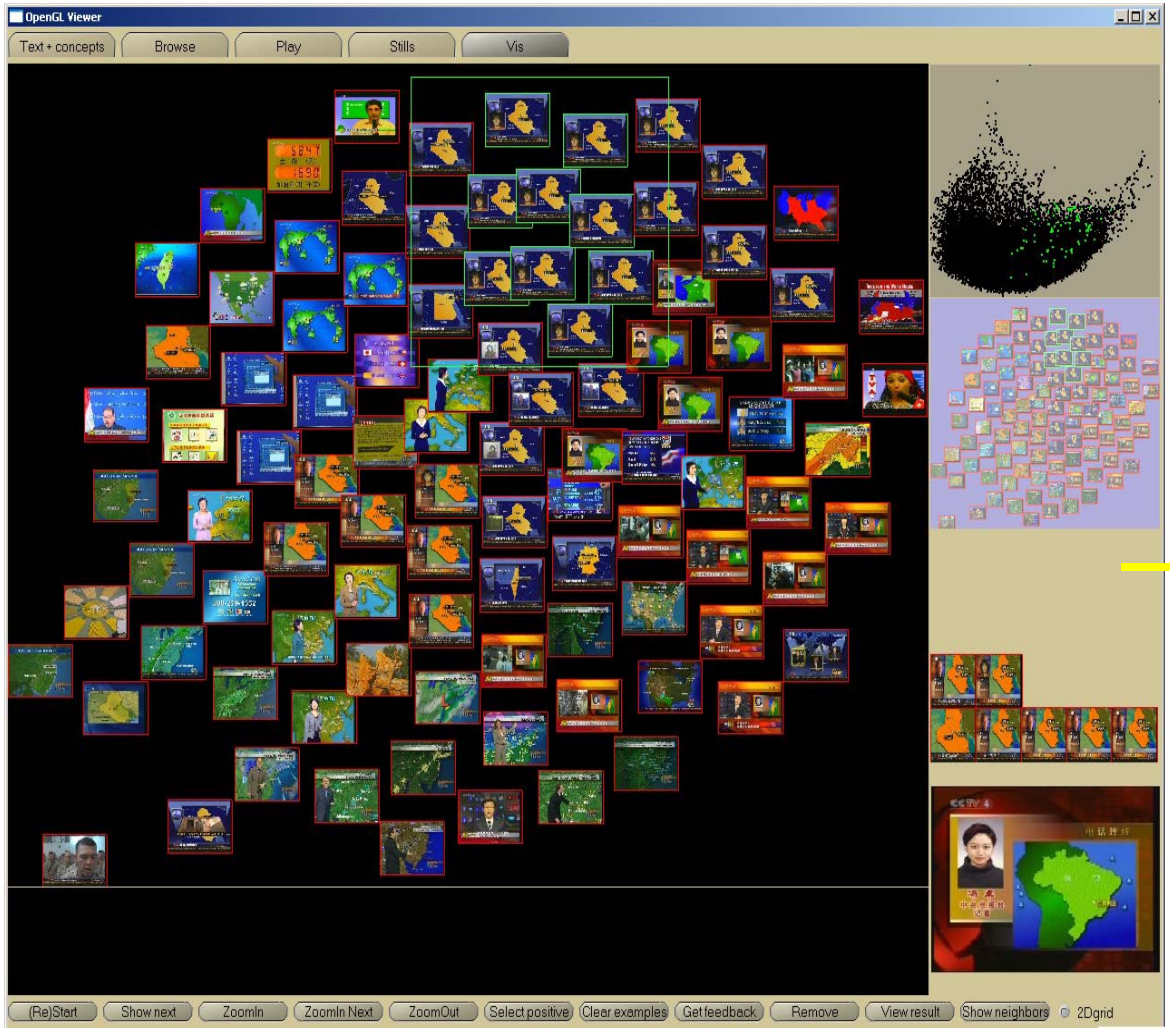
- Analyze the context of the current shot



The GalaxyBrowser

- What
- Why
- How
- Why not?
- Conclusion

➢ Pure similarity based browsing

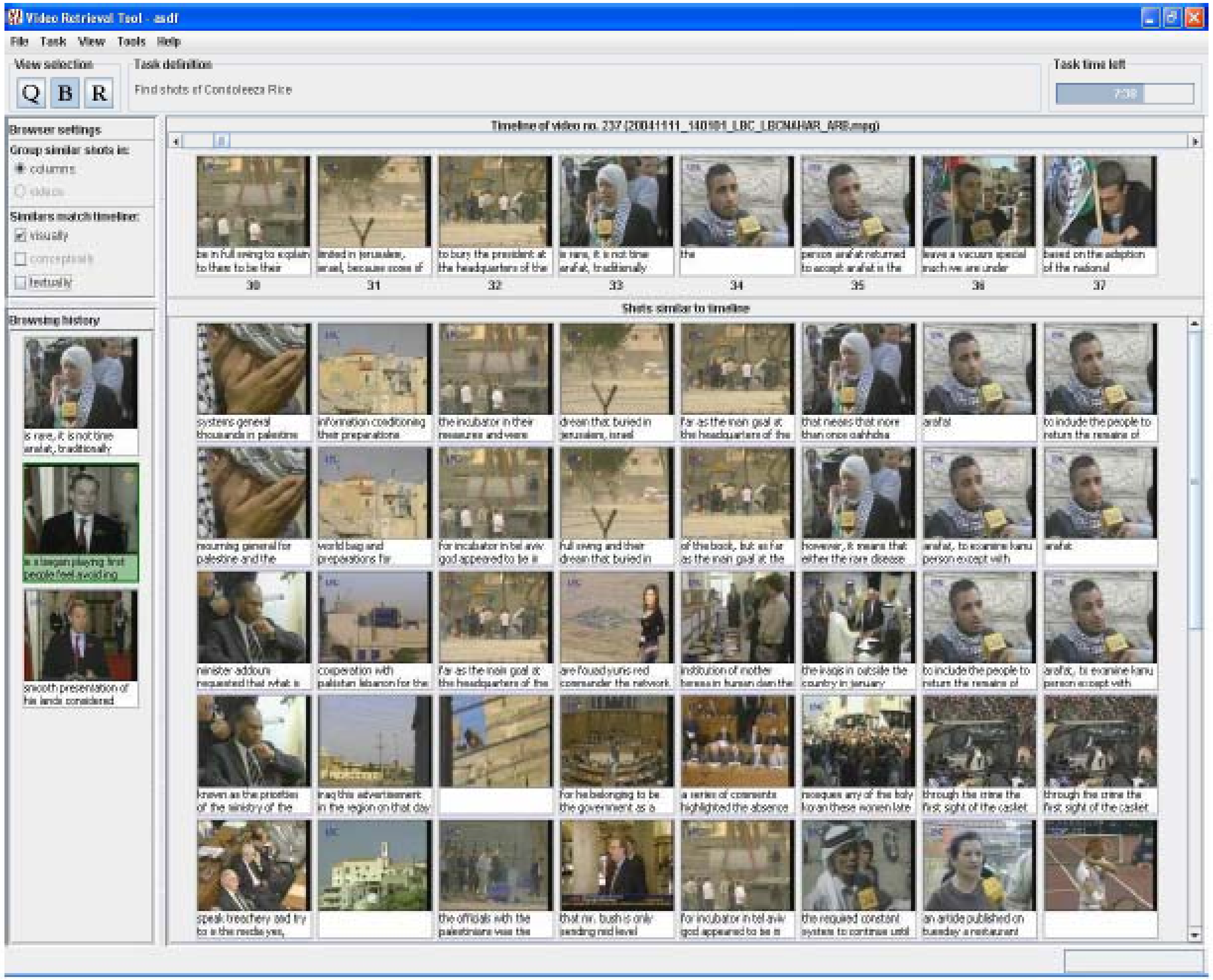


Induced by similarity

Cluster-temporal browsing

- What
- Why
- How
- Why not?
- Conclusion

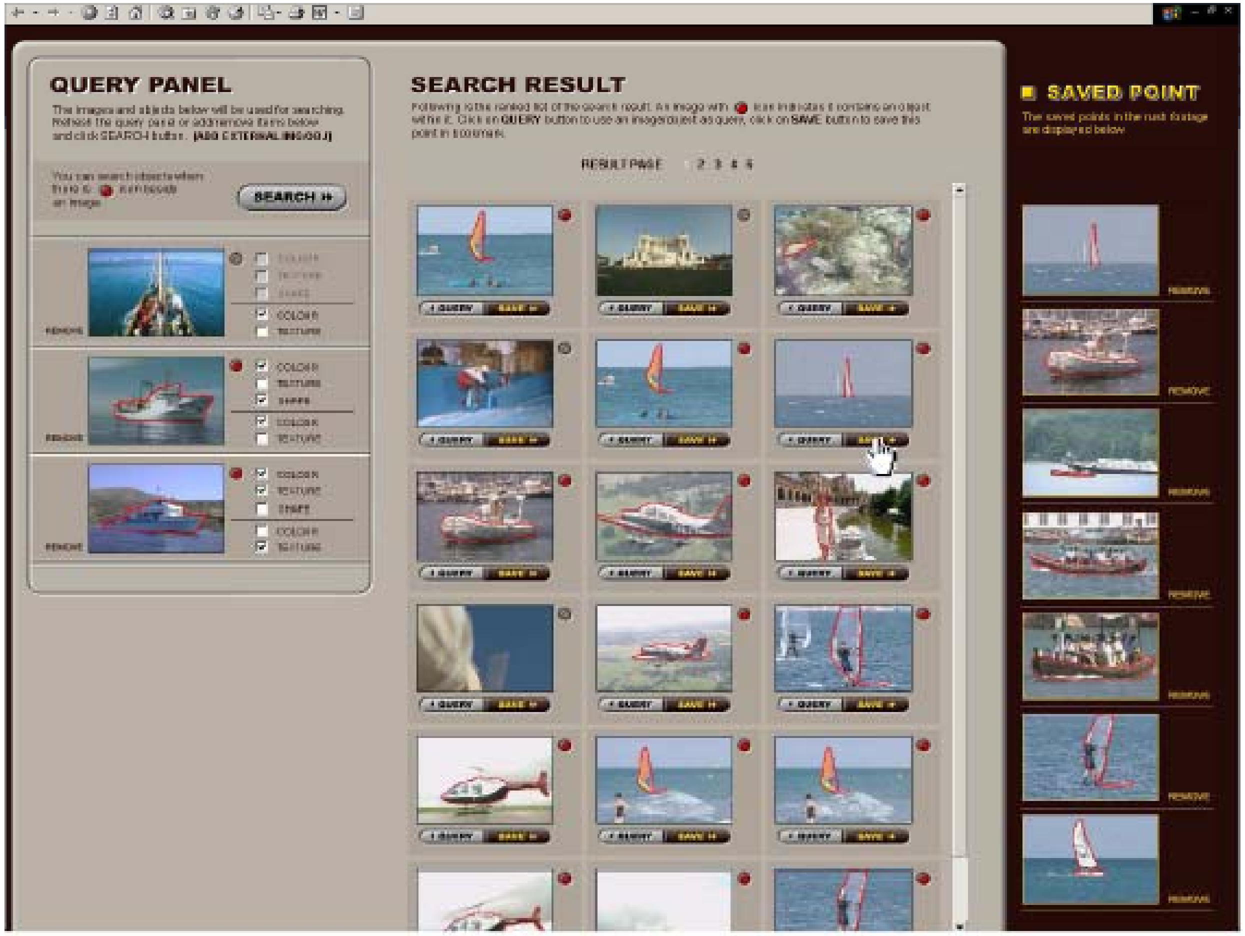
➤ Using that result are typically similar/close in time



Físchlár

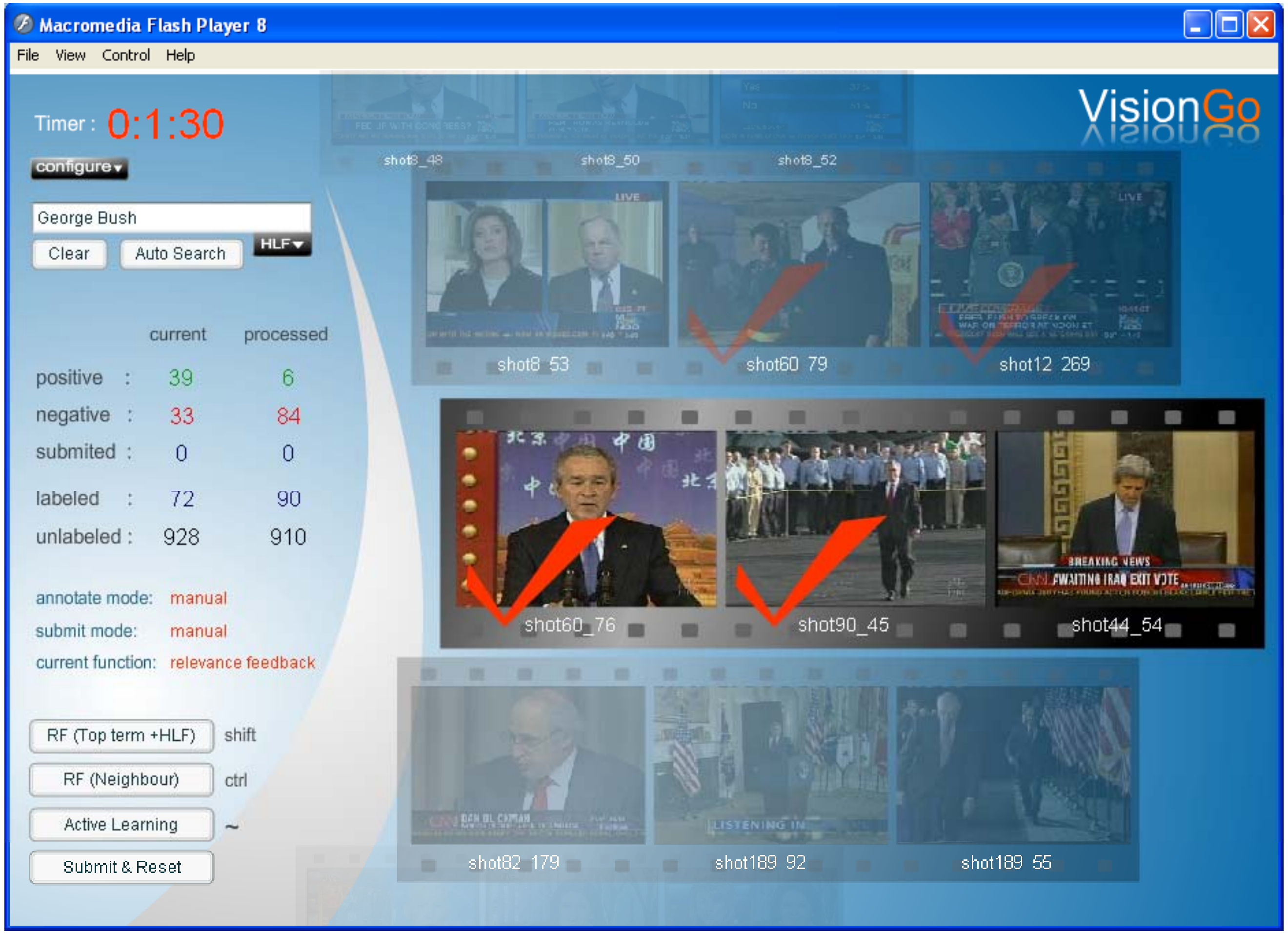
- What
- Why
- How
- Why not?
- Conclusion

➢ Optimized for use by "real" users



VisionGo

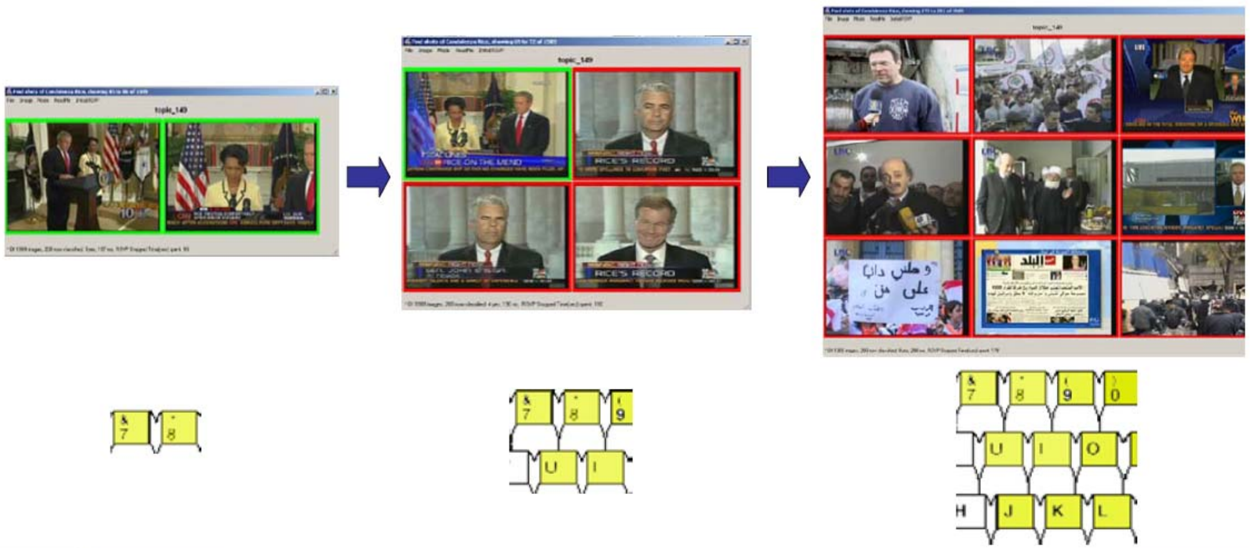
➢ Extremely fast and efficient



Extreme video retrieval

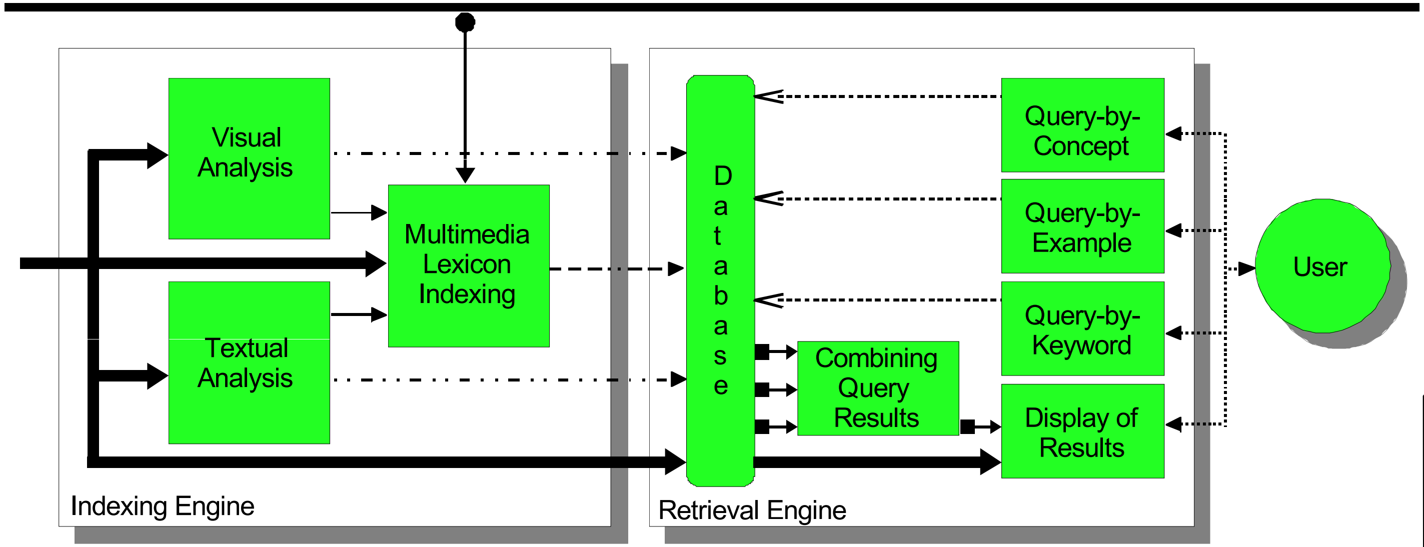
- What
- Why
- How
- Why not?
- Conclusion

- **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



Lessons learned

- What
- Why
- How
- Why not?
- Conclusion



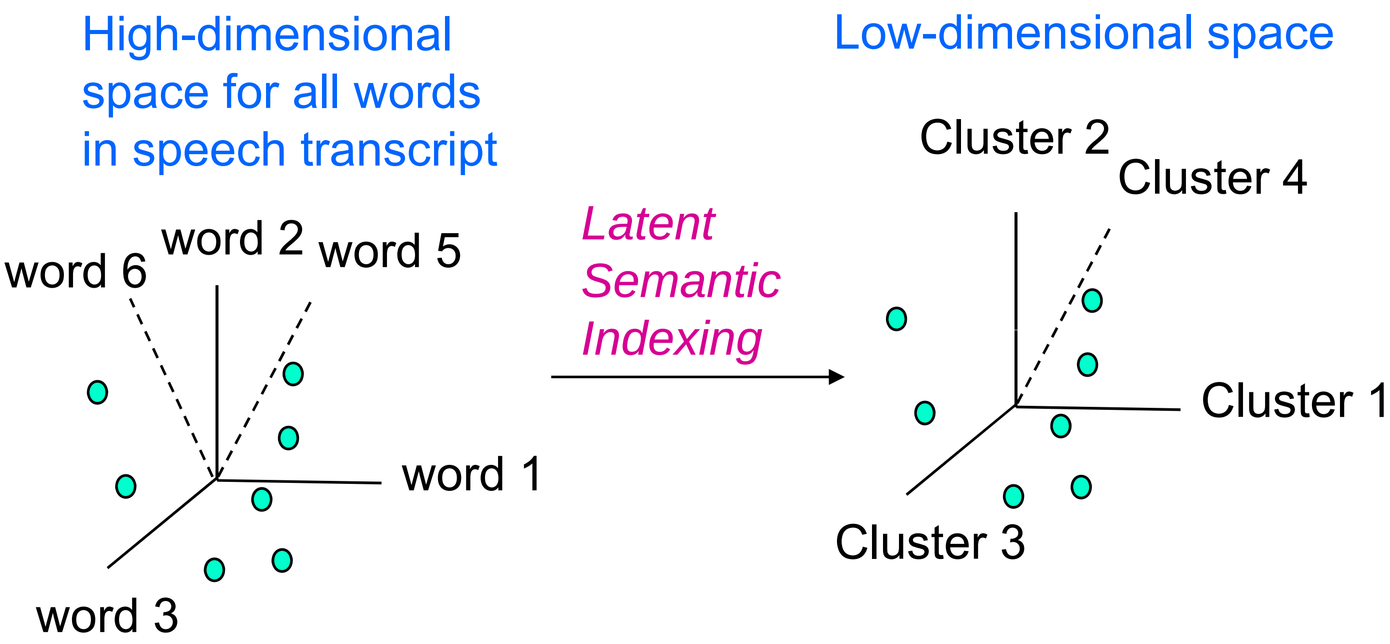
- Do not ignore text analysis.
  - ✓ Provides a valuable baseline
- Do not ignore the interface
  - ✓ You can have too much of a good thing (and too few…)
- Do not ignore evaluation of interactive retrieval
  - ✓ Preferably using common benchmarks

- What is the influence of an increasing lexicon?
  - ✓ Well we need a video search engine first…

Textual analysis

Classical Techniques

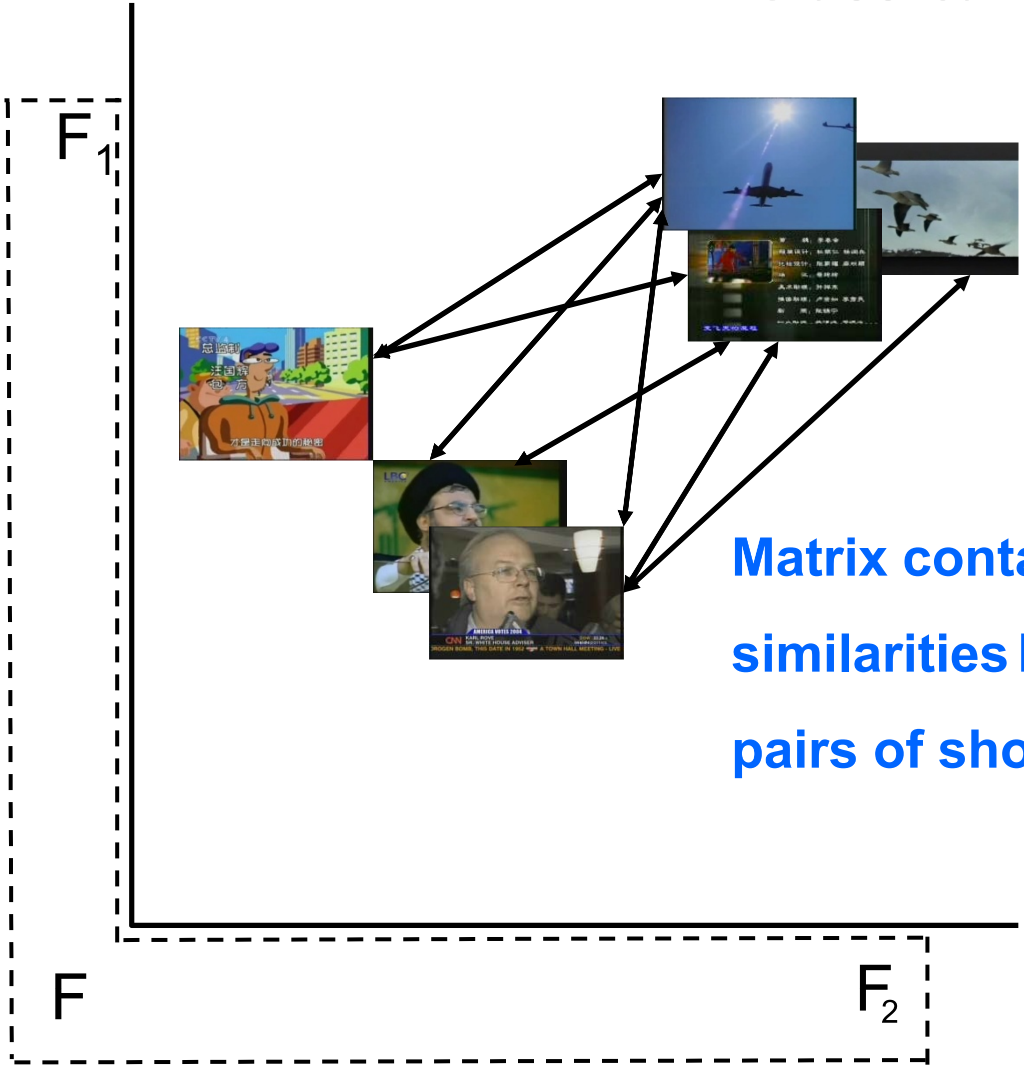
- What
- Why
- How
- Why not?
- Conclusion



Visual analysis

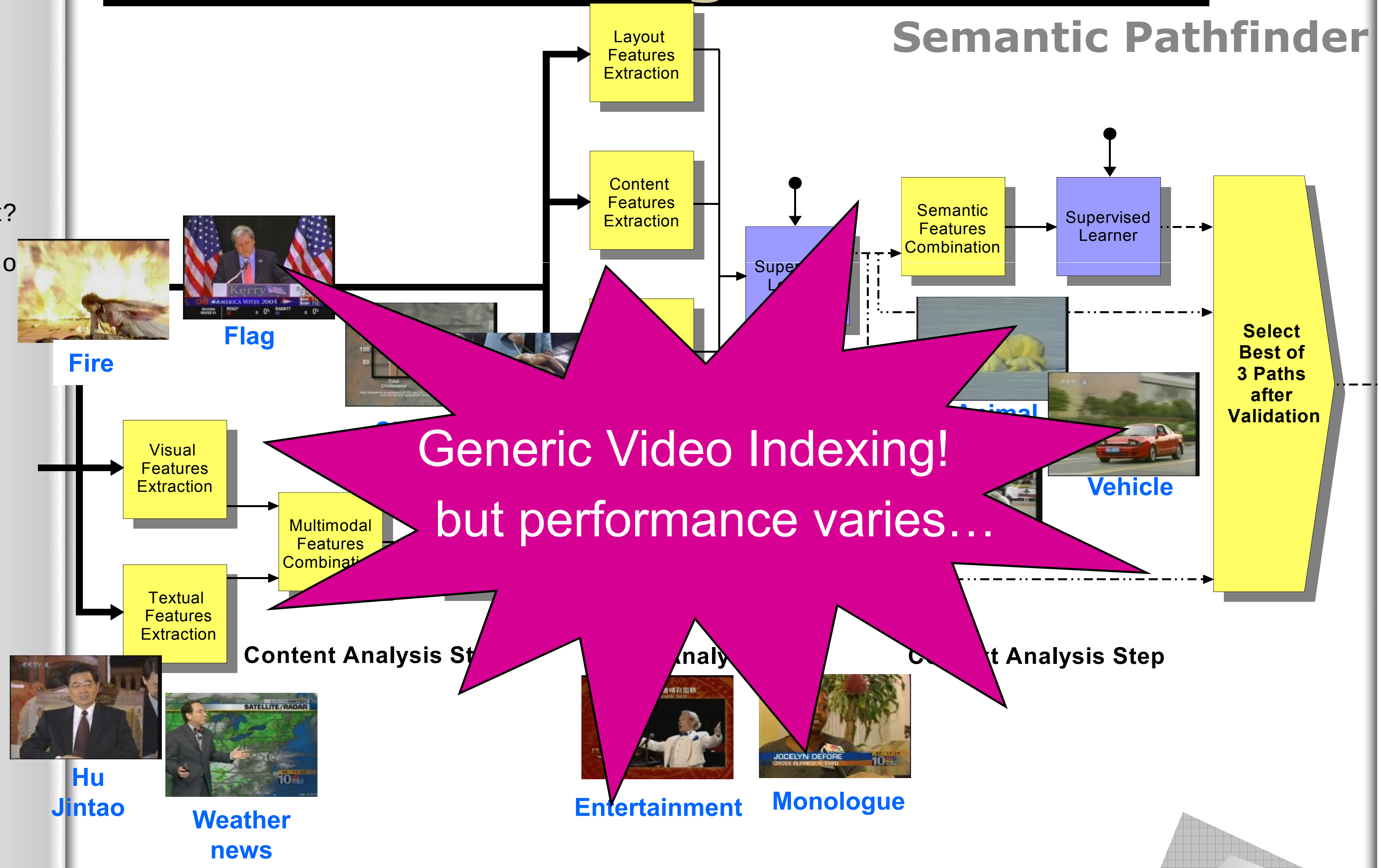
- What
- Why
- How
- Why not?
- Conclusion

Classical Techniques



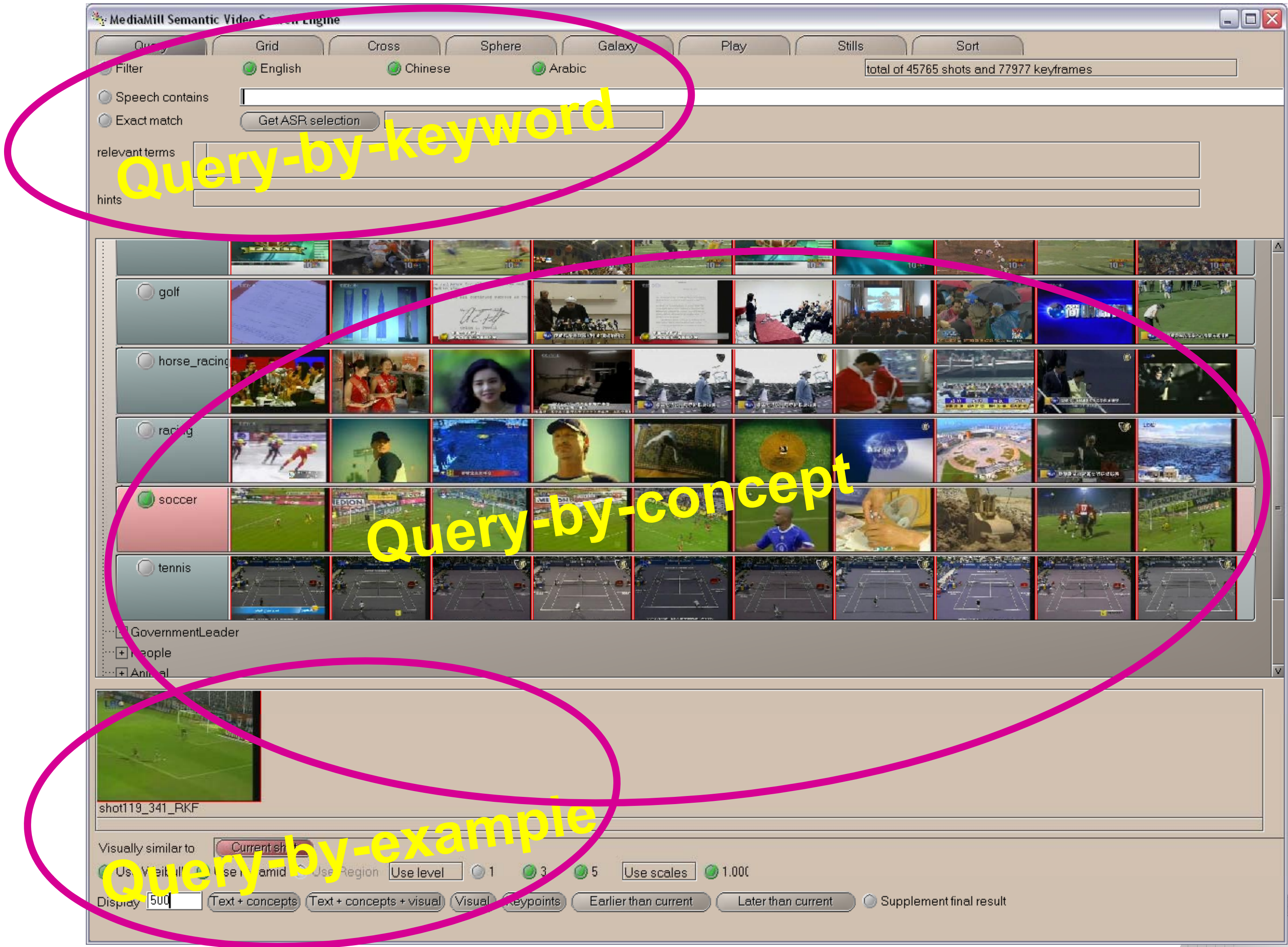
Matrix containing all similarities between pairs of shots

Lexicon indexing



Query selection

- What
- Why
- How
- Why not?
- Conclusion

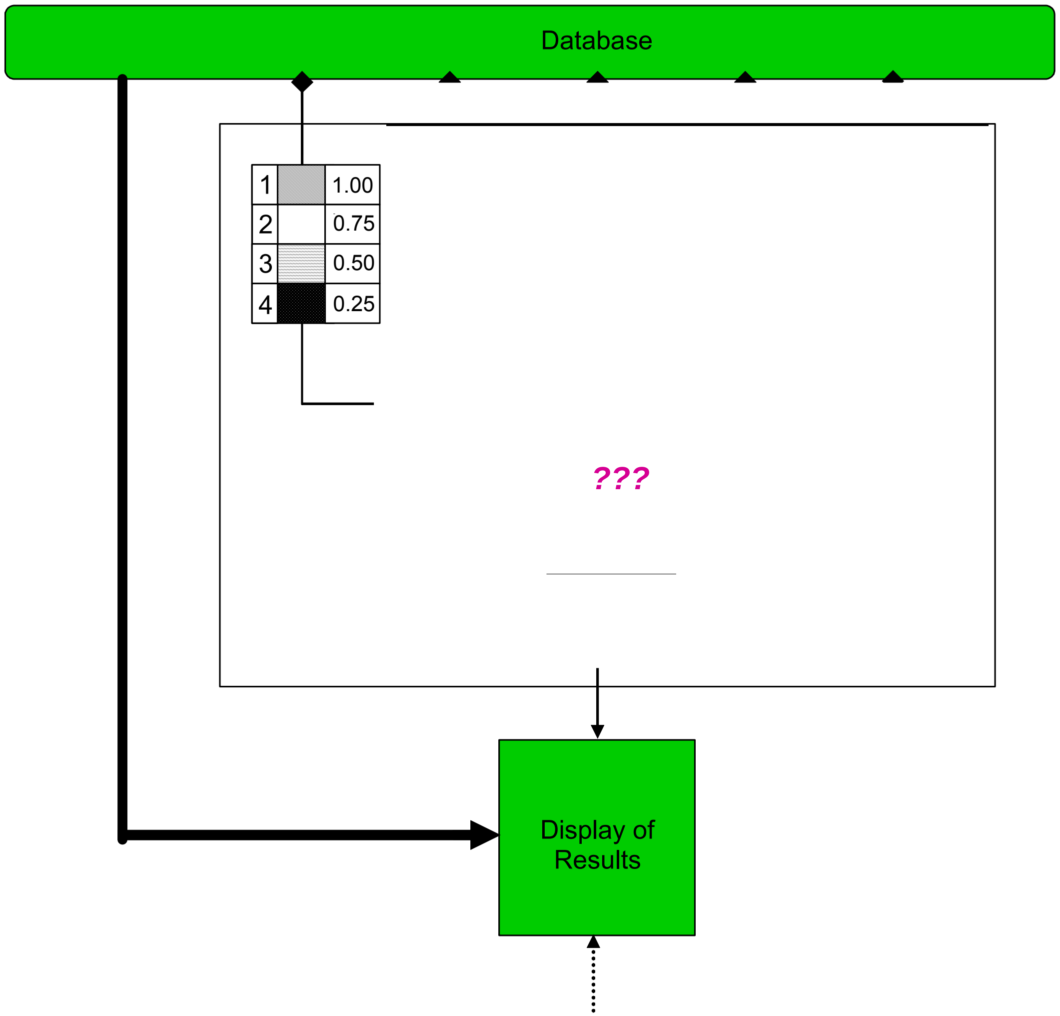


… yields a ranking of the data

Combining query results

Classical Techniques

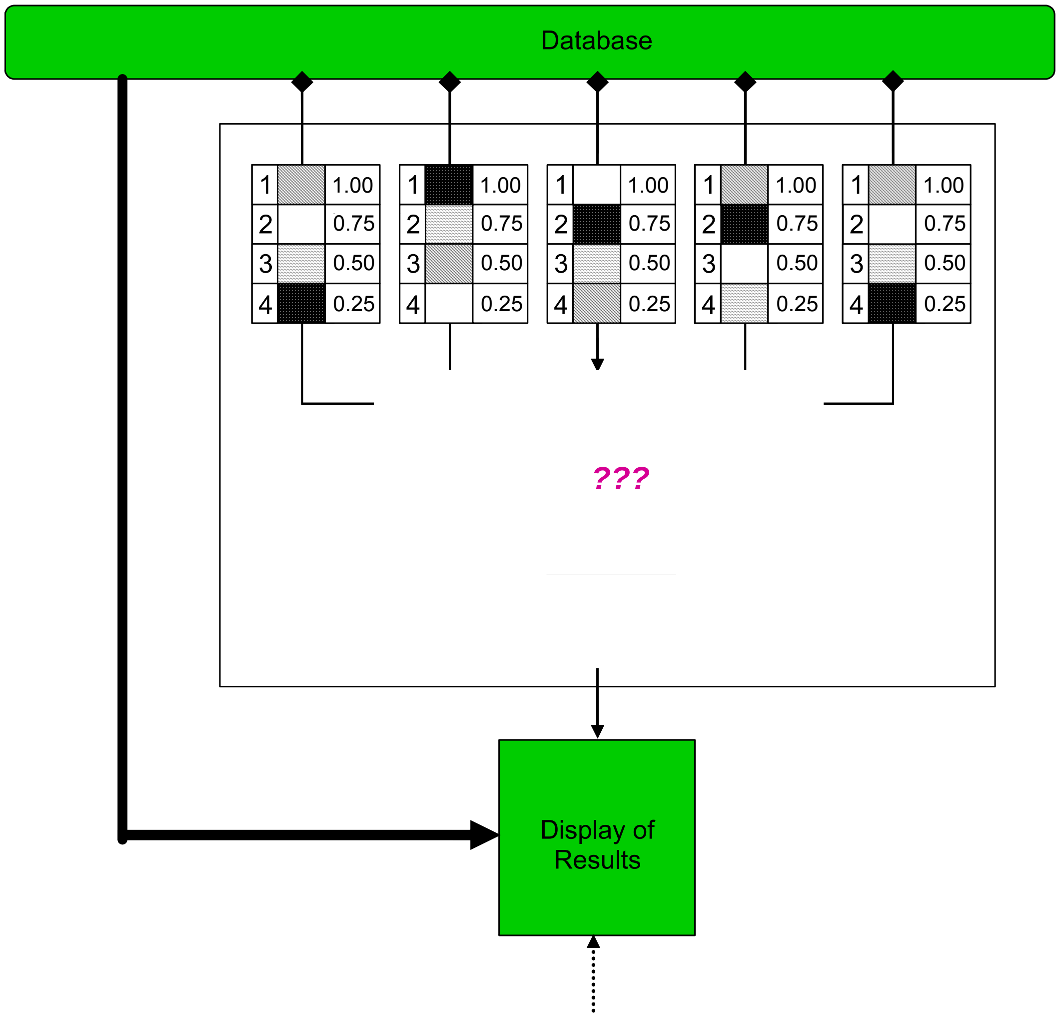
- What
- Why
- How
- Why not?
- Conclusion



Combining query results

Classical Techniques

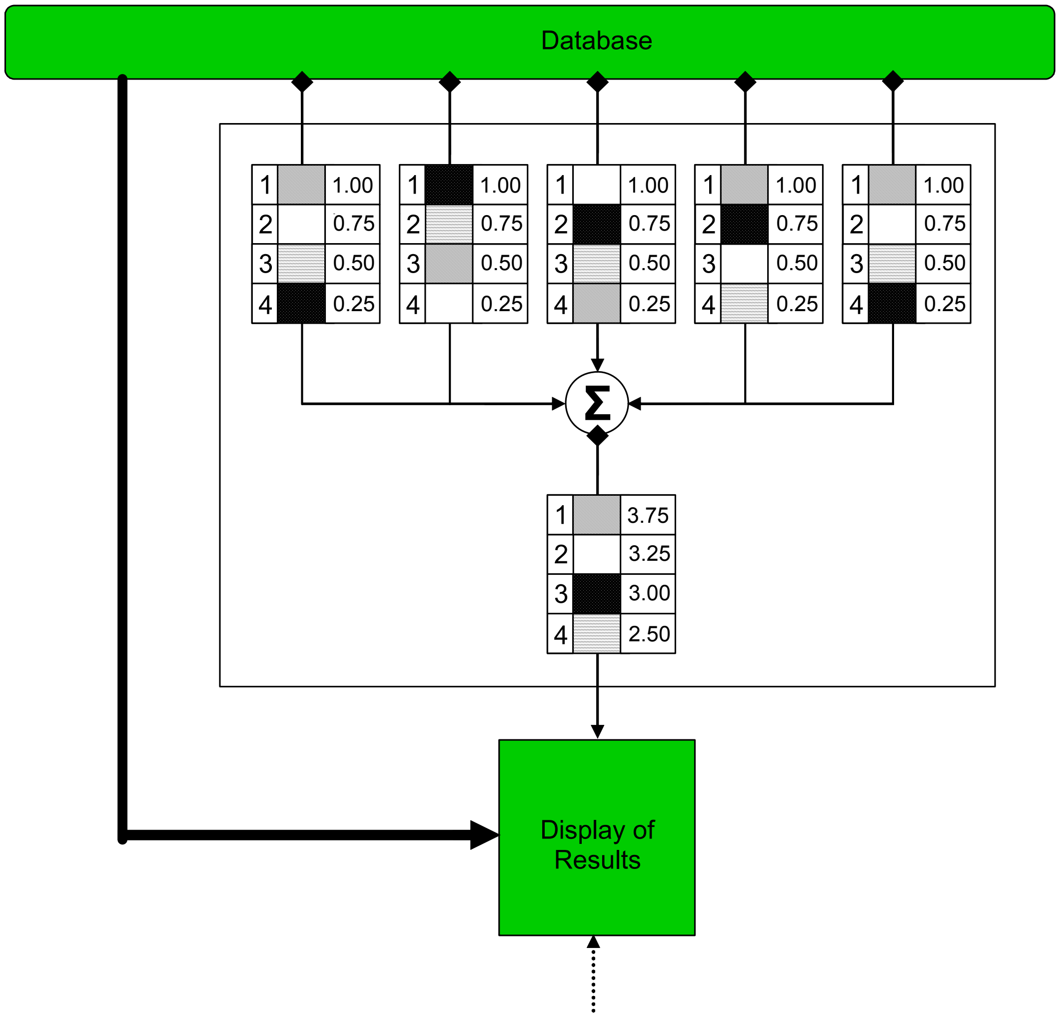
- What
- Why
- How
- Why not?
- Conclusion



- What
- Why
- How
- Why not?
- Conclusion

Combining query results

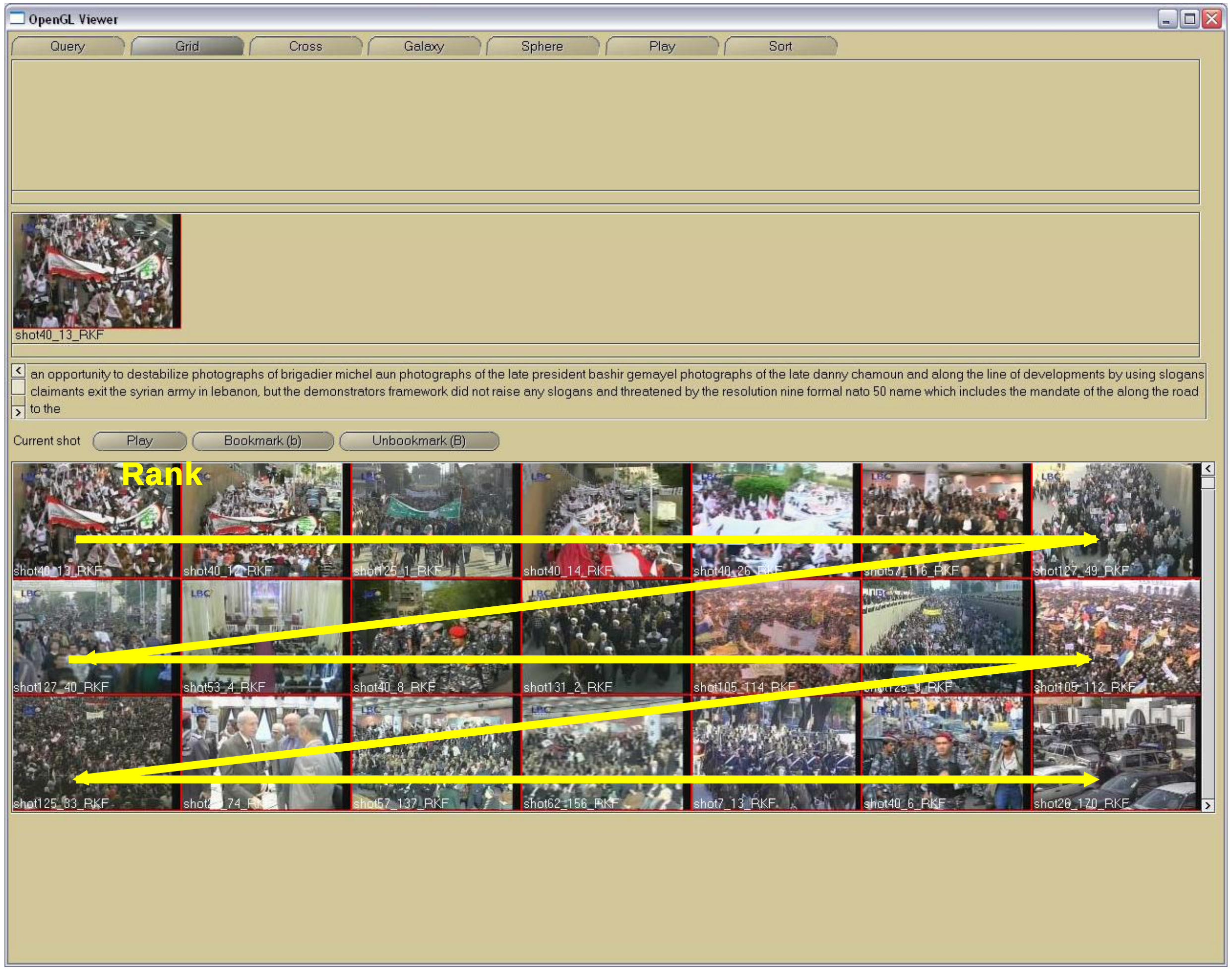
Classical Techniques



Display of results

Classical Techniques

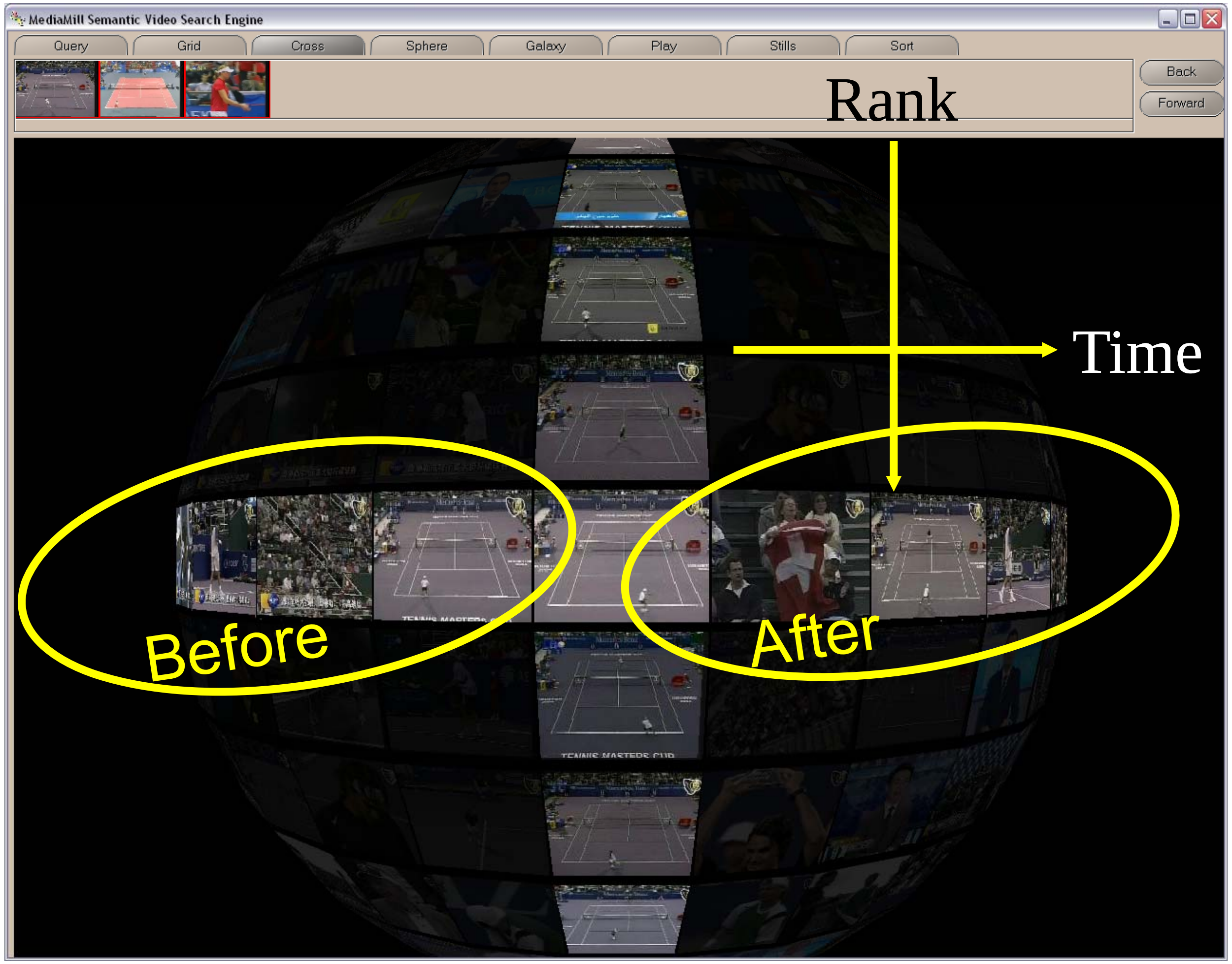
- What
- Why
- How
- Why not?
- Conclusion



Display of results

- What
- Why
- How
- Why not?
- Conclusion

CrossBrowser

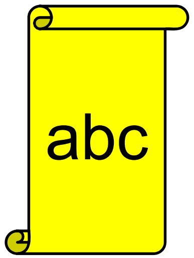


NIST TRECVID benchmark

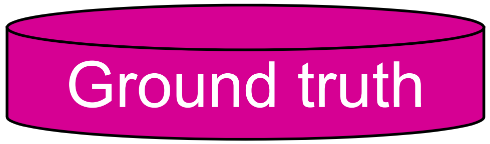
anno 2001

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

Data set



Speech transcript



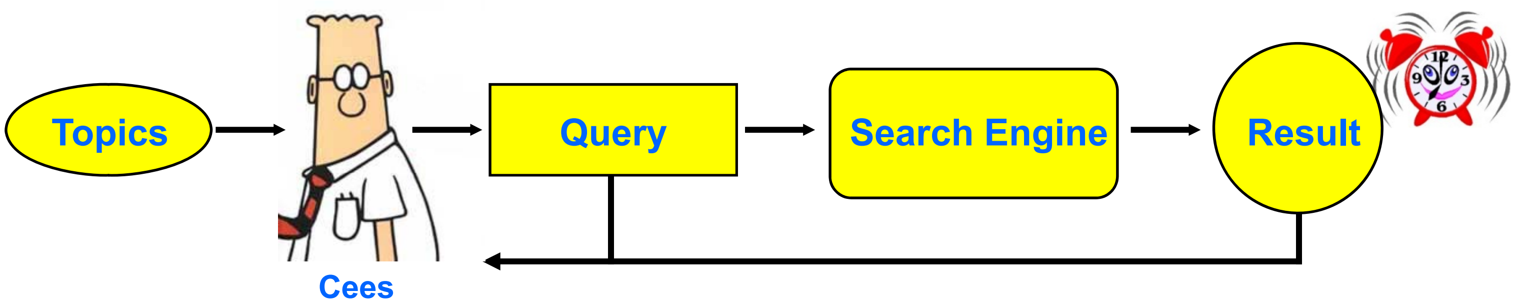
- Large international field of participants

- Currently the de facto standard for evaluation

http://trecvid.nist.gov/

TRECVID interactive search task

- TRECVID interactive retrieval procedure:



- ✓ NOTE: topics unknown at time of lexicon creation!
- ✓ NOTE: user had training set knowledge of concept detectors

TRECVID search topics

- What
- Why
- How
- Why not?
- Conclusion

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

Find shots of a meeting with a large table and people

Find shots of one or more helicopters in flight.

Find shots of a goal being made in a soccer match

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.

Experimental Setup

- What
- Why
- How
- Why not?
- Conclusion

- 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)

Learned lexicon of 32 concepts

- What
- Why
- How
- Why not?
- Conclusion


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

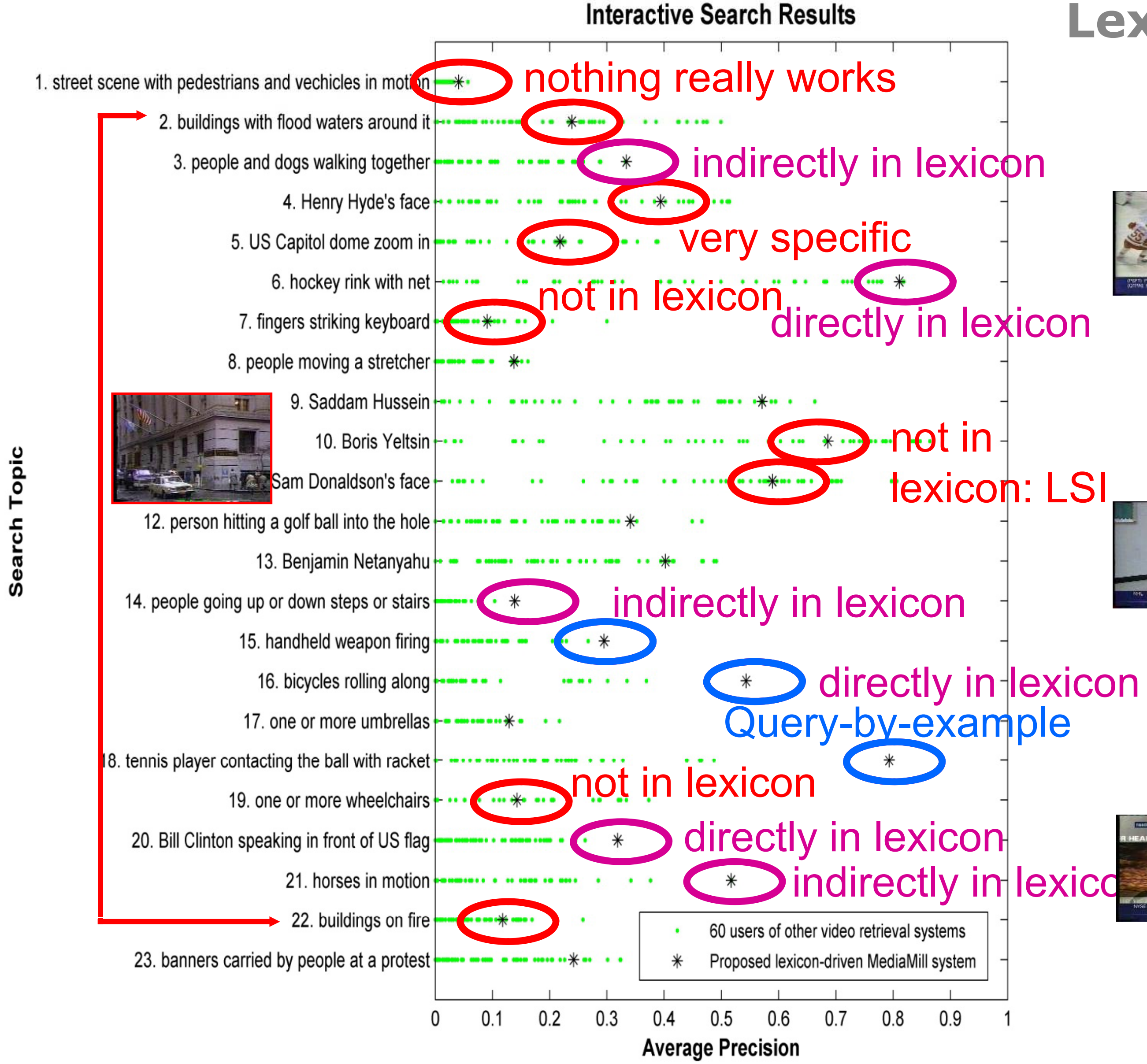

Madeleine Albright


Soccer


Bicycle


Weather News

Experiment 1



Lexicon = 32 concepts

The Good

(in)directly in lexicon

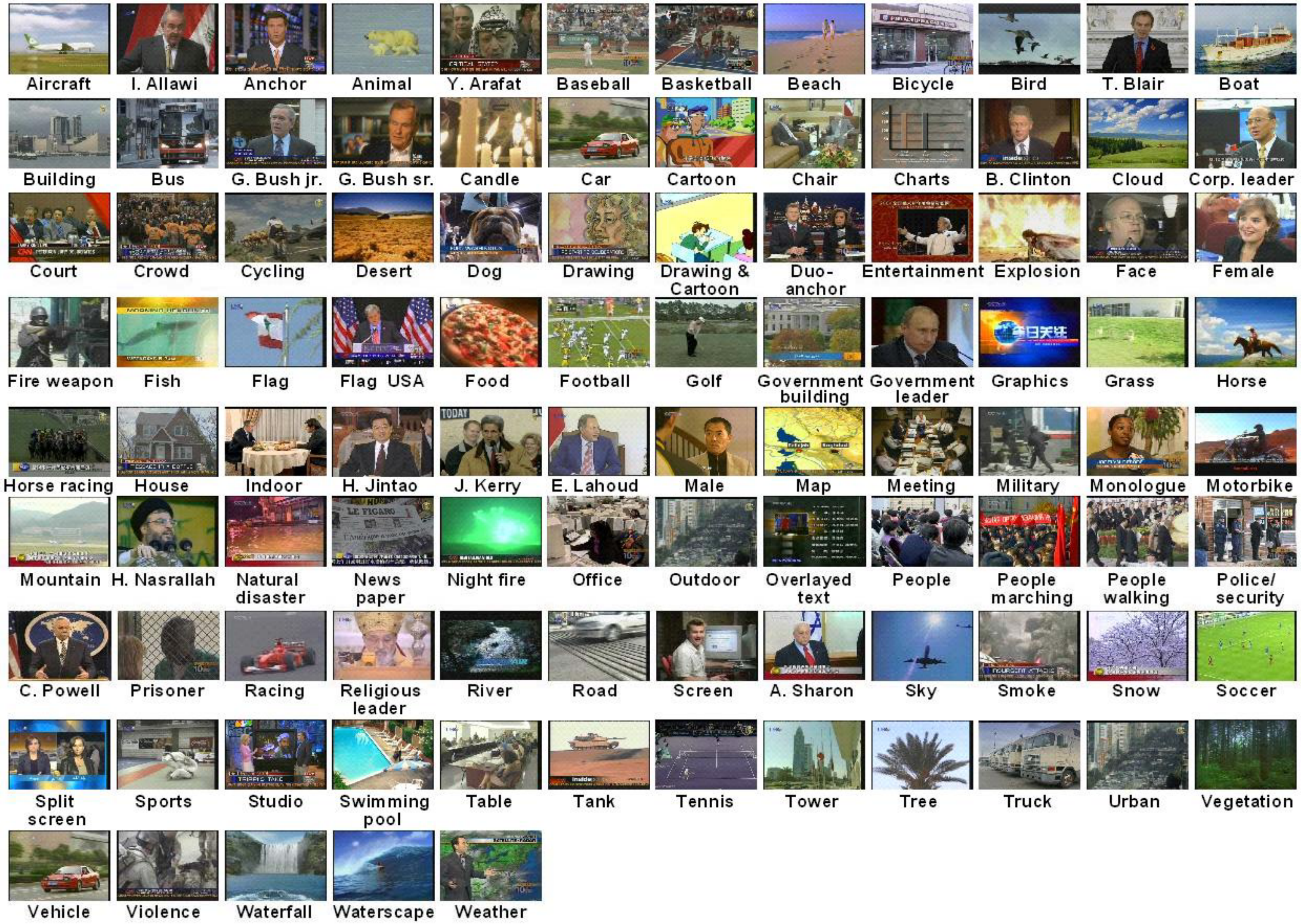
The Bad

not in lexicon

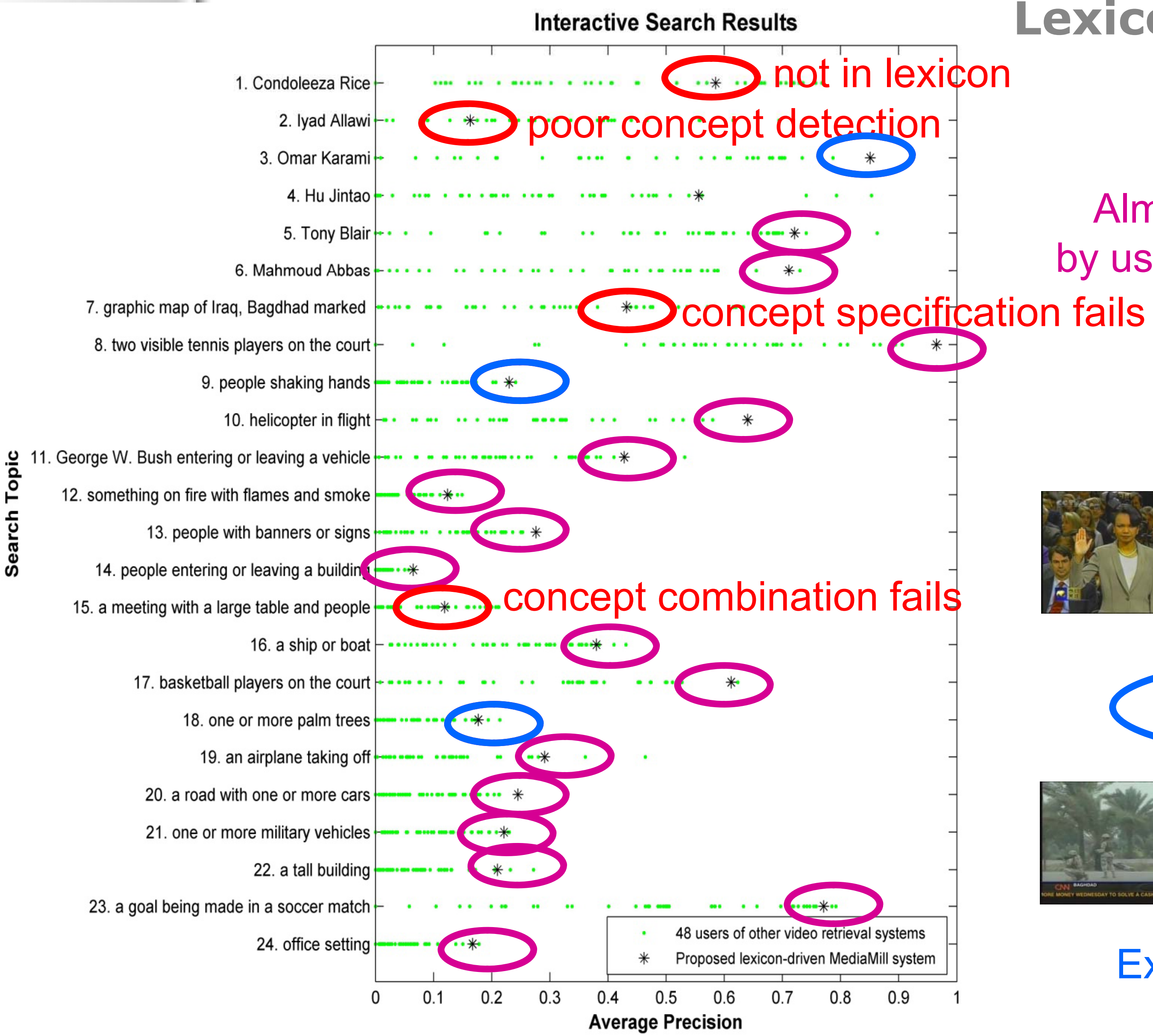
The Ugly

exploit TV repetition

Learned lexicon of 101 concepts



Experiment 2



Lexicon = 101 concepts

The Good

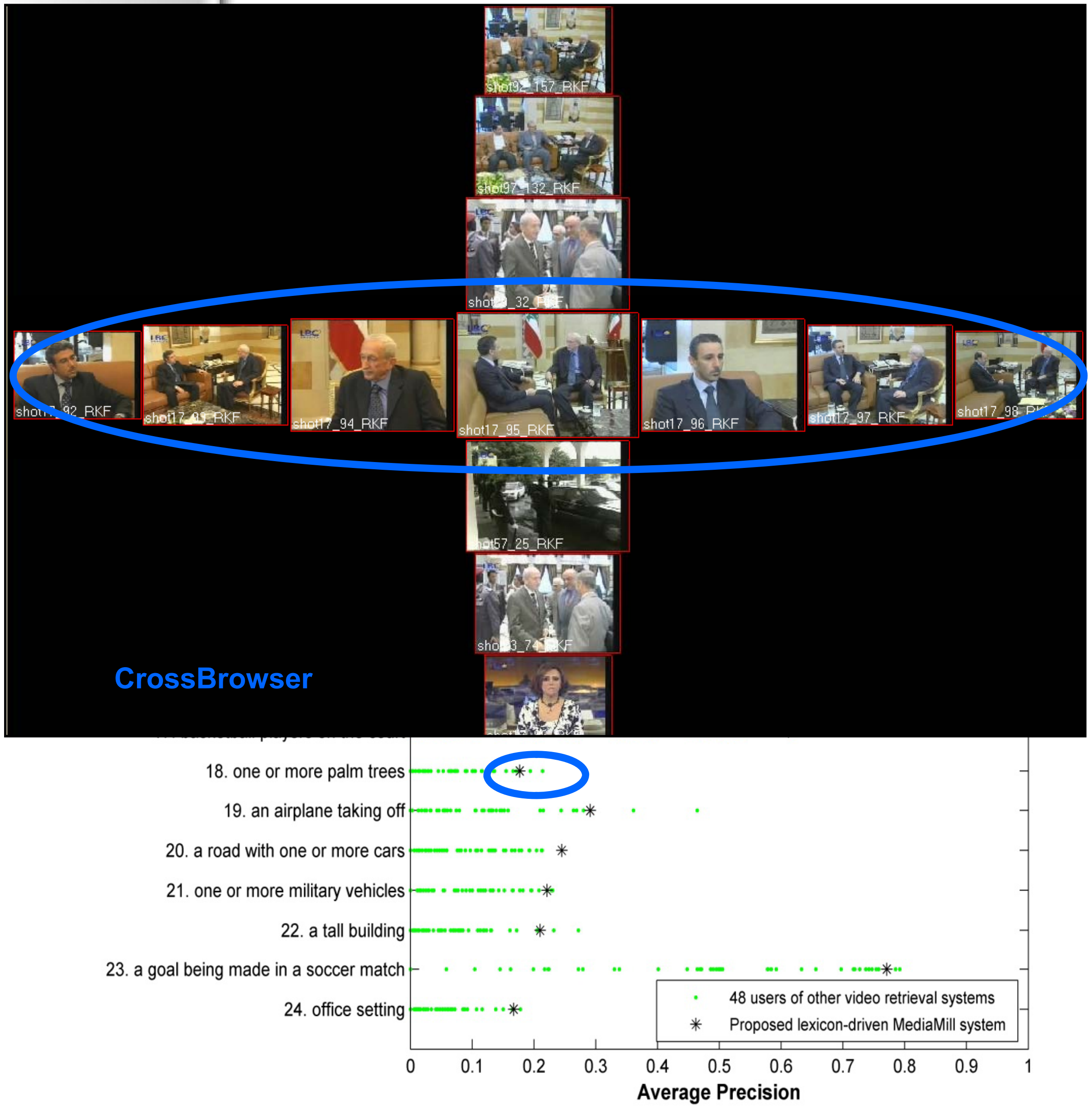
Almost all topics solvable
by using concept lexicon only!

The Bad

The Beautiful

Exploit common sense!

Experiment 2



Lexicon = 101 concepts

The Good

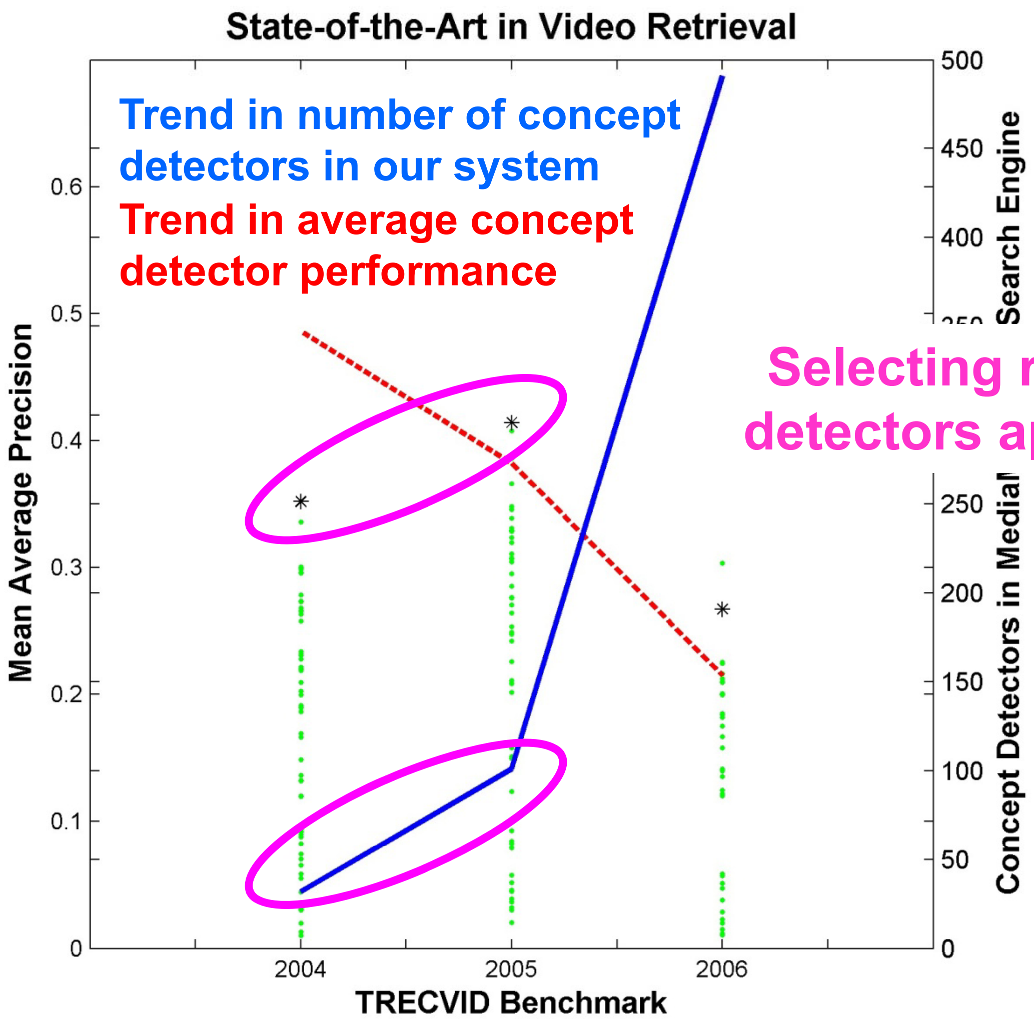
Almost all topics solvable
by using concept lexicon only!

The Bad

The Beautiful

Exploit common sense!

MediaMill @ TRECVID



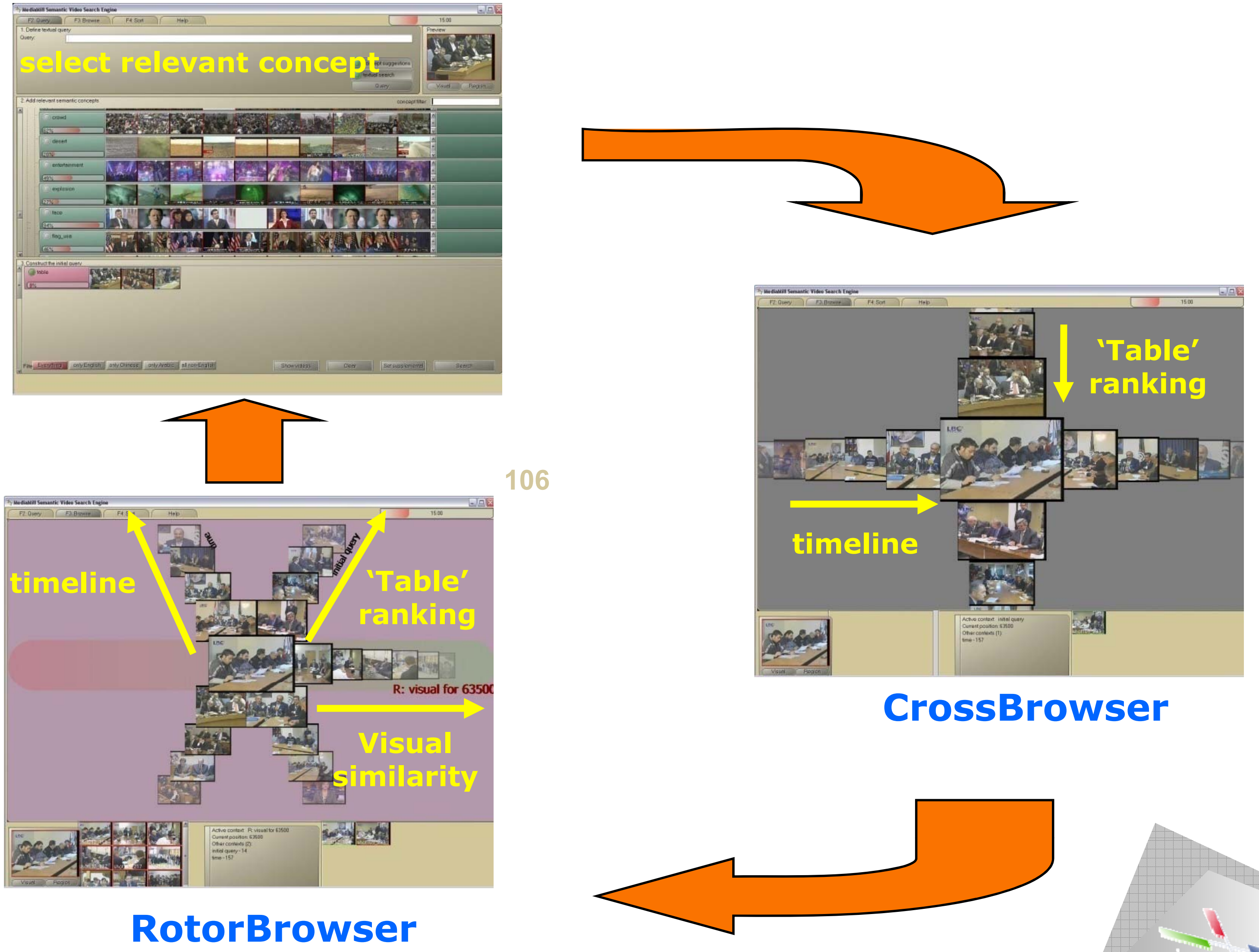
Selecting robust and relevant detectors appears to be difficult

- 143 performance evaluations of other systems

* MediaMill Semantic Video Search Engine

Demo time!

- What
- Why
- How
- Why not?
- Conclusion



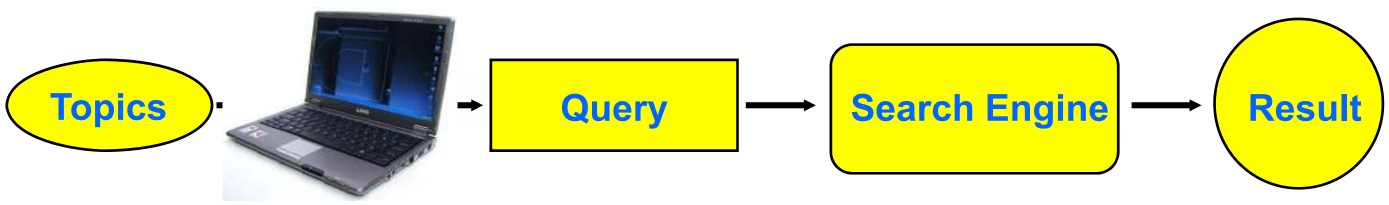
Conclusions I

- Interactive Video Retrieval is an interplay of:
  - ✓ Various query selection methods *(keyword, example, concept)*
  - ✓ An advanced display of results
  - ✓ A goal-oriented human user
- What matters most?
  - ✓ A large lexicon of semantic concepts
  - ✓ With only 32 concepts, we already outperform state-of-the-art systems in 7 out of 23 random queries *(and overall best)*
  - ✓ With only 101 concepts, we solve 17 out of 24 random queries with highly competitive accuracy *(and overall best)*
  - ✓ How many queries can we solve with 1001 concepts?
- Unanswered questions
  - ✓ How to combine semantic concepts?
  - ✓ How to equip machines with common sense?
  - ✓ How to include user experience?
  - ✓ How to visualize semantic space?
  - ✓ Is an ontology the answer to our questions?

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
- Lets 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.*

TRECVID automatic search task



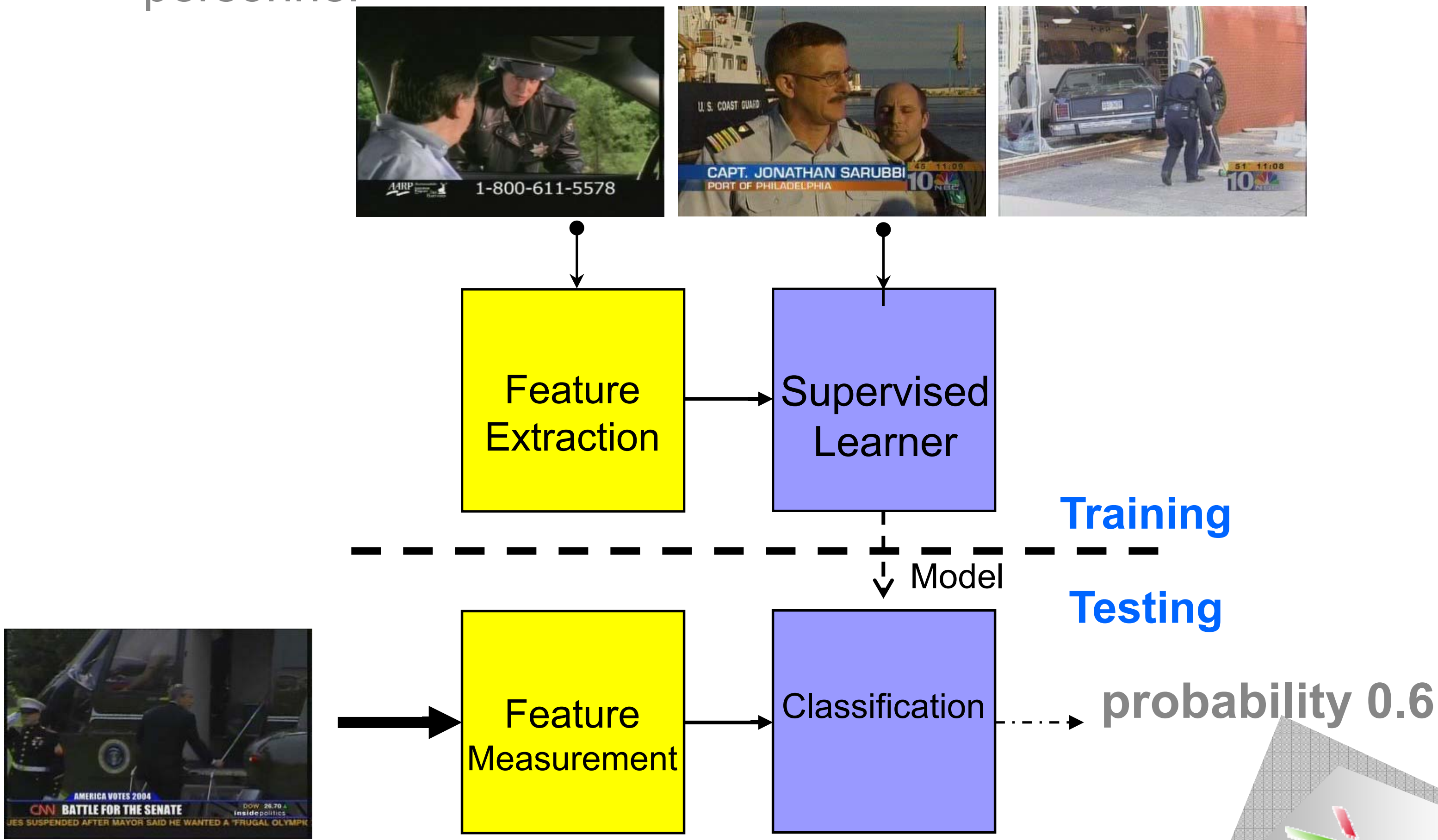
- Automatically solve search topic
- Return 1,000 ranked shot-based results
- Evaluate using Average Precision

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

- What
- Why
- How
- Why not?
- Conclusion

Recap: a simple concept detector

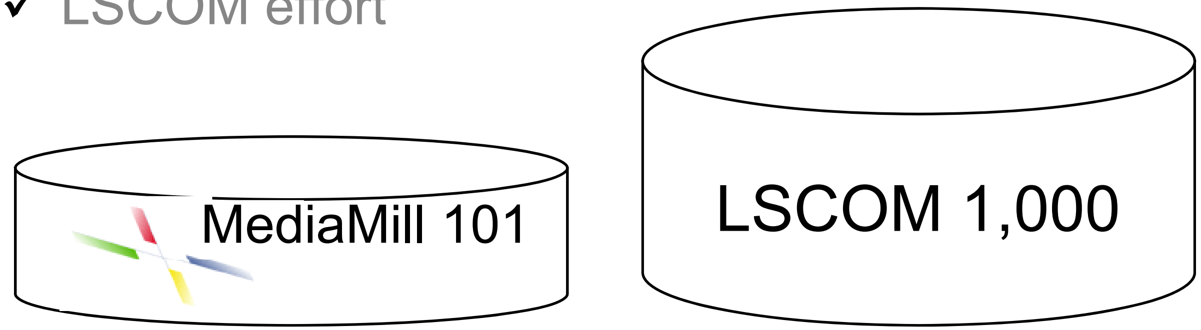
- MM078-Police/Security Personnel
  - ✓ Shots depicting law enforcement or private security agency personnel



References:
Naphade et al, IEEE Multimedia 2006

How to obtain concept detectors?

- Large lexicons of concept detectors are available
  - ✓ MediaMilll Challenge
  - ✓ LSCOM effort



-Annotations
-Features
-Detectors
-Performance

-Annotations

363 detectors

MediaMill: **http://www.mediamill.nl/challenge/**
LSCOM: **http://www.ee.columbia.edu/ln/dvmm/lscom/**

Influence of lexicon size

TRECVID 2005

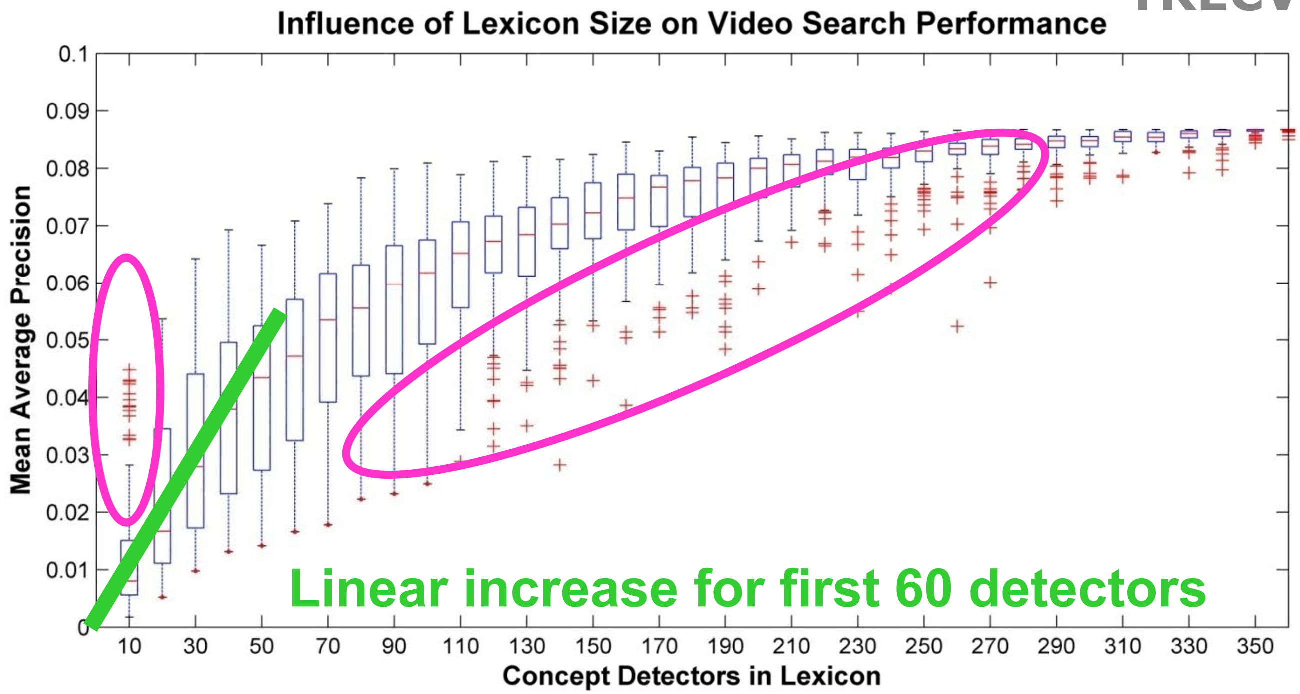
- Lexicon = 363 machine learned concept detectors

- 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

Influence of lexicon size

TRECVID 2005



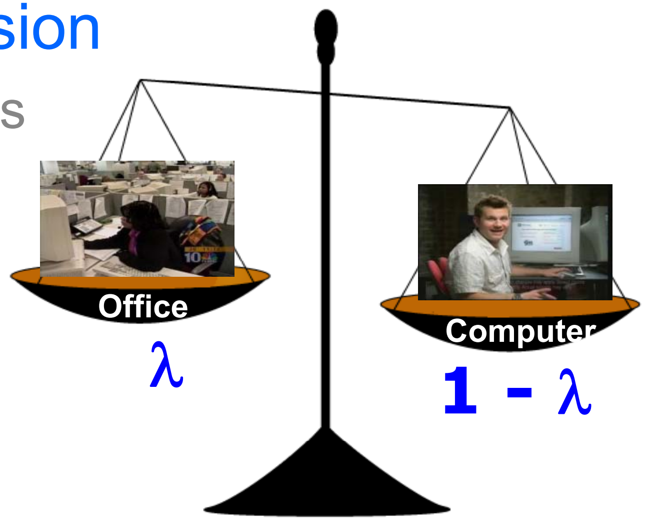
- Size matters
  - ✓ Lexicon of 150 detectors comes close to maximum performance
- Some detectors perform well for specific topics
  - ✓ Tennis game detector for “find two visible tennis players”
- Substantial number of detectors not accurate enough yet
  - ✓ Only small increase when more than 70 detectors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

- What
- Why
- How
- Why not?
- Conclusion

Influence of detector combination

TRECVID 2005

- Experiment: pair-wise oracle fusion
  - ✓ Improvement for 20 out of 24 topics
  - ✓ Increase per topic as high as 89%
  - ✓ Overall increase 10%



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

Best | 2nd Best


+


Maps | 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)

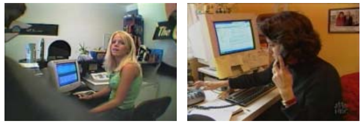

+

?

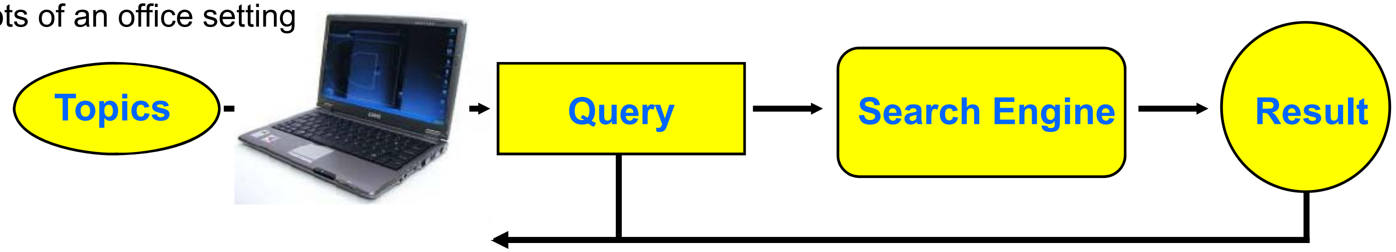
rocket propelled grenades | Iyad Allawi

Problem statement

- What
- Why
- How
- Why not?
- Conclusion



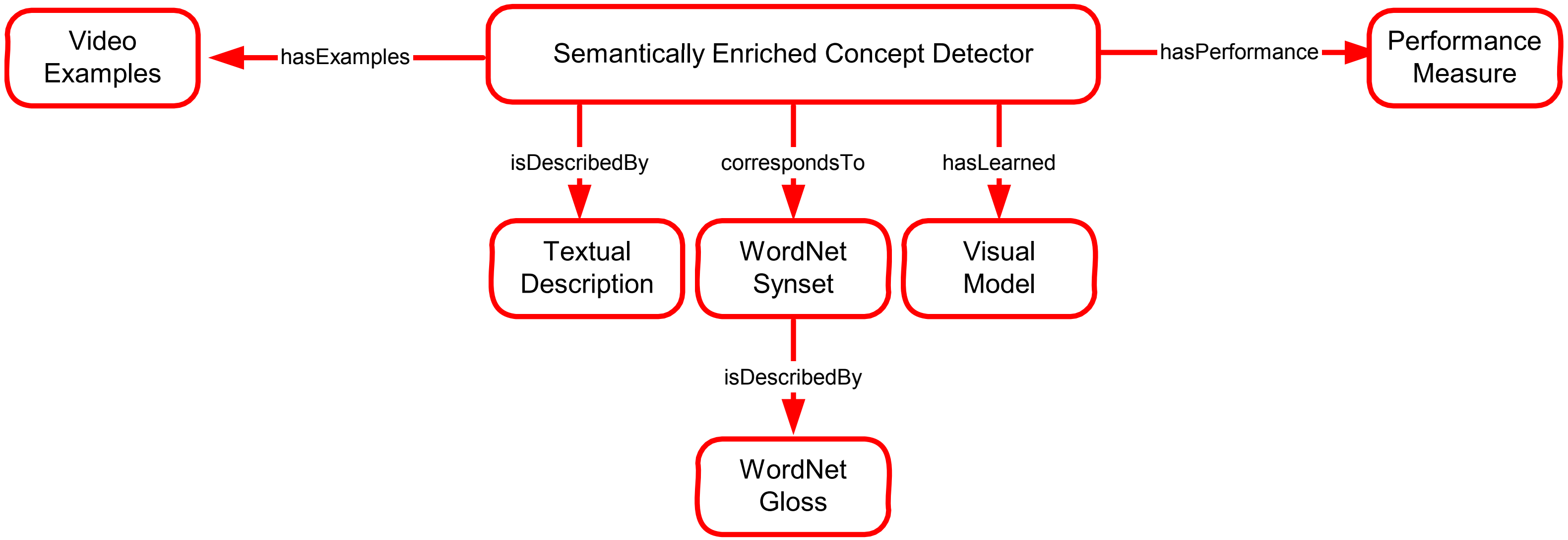
Find shots of an office setting



➢ How to translate query topic to semantics?

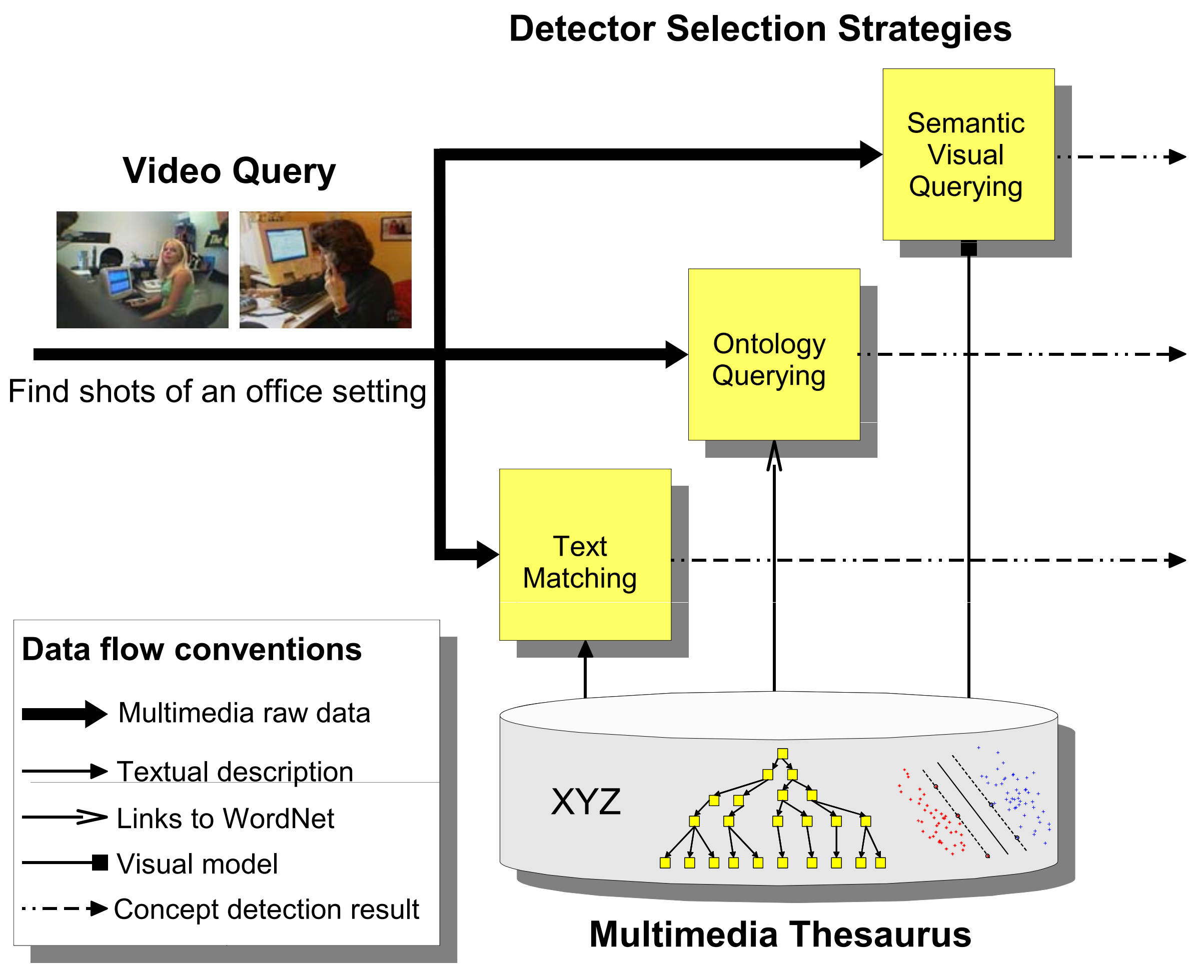
Adding semantics to detectors

- Enrich concept detectors by adding:
  - ✓ Text
  - ✓ WordNet links
  - ✓ Vsual models



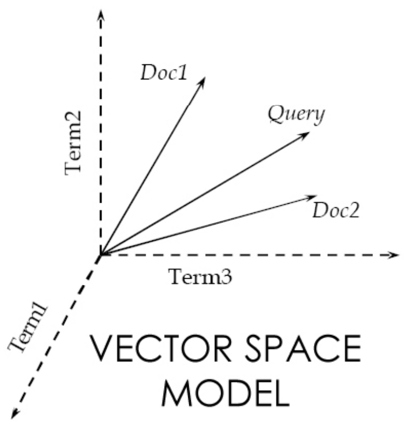
Detector selection strategies

- What
- Why
- How
- Why not?
- Conclusion



Text matching

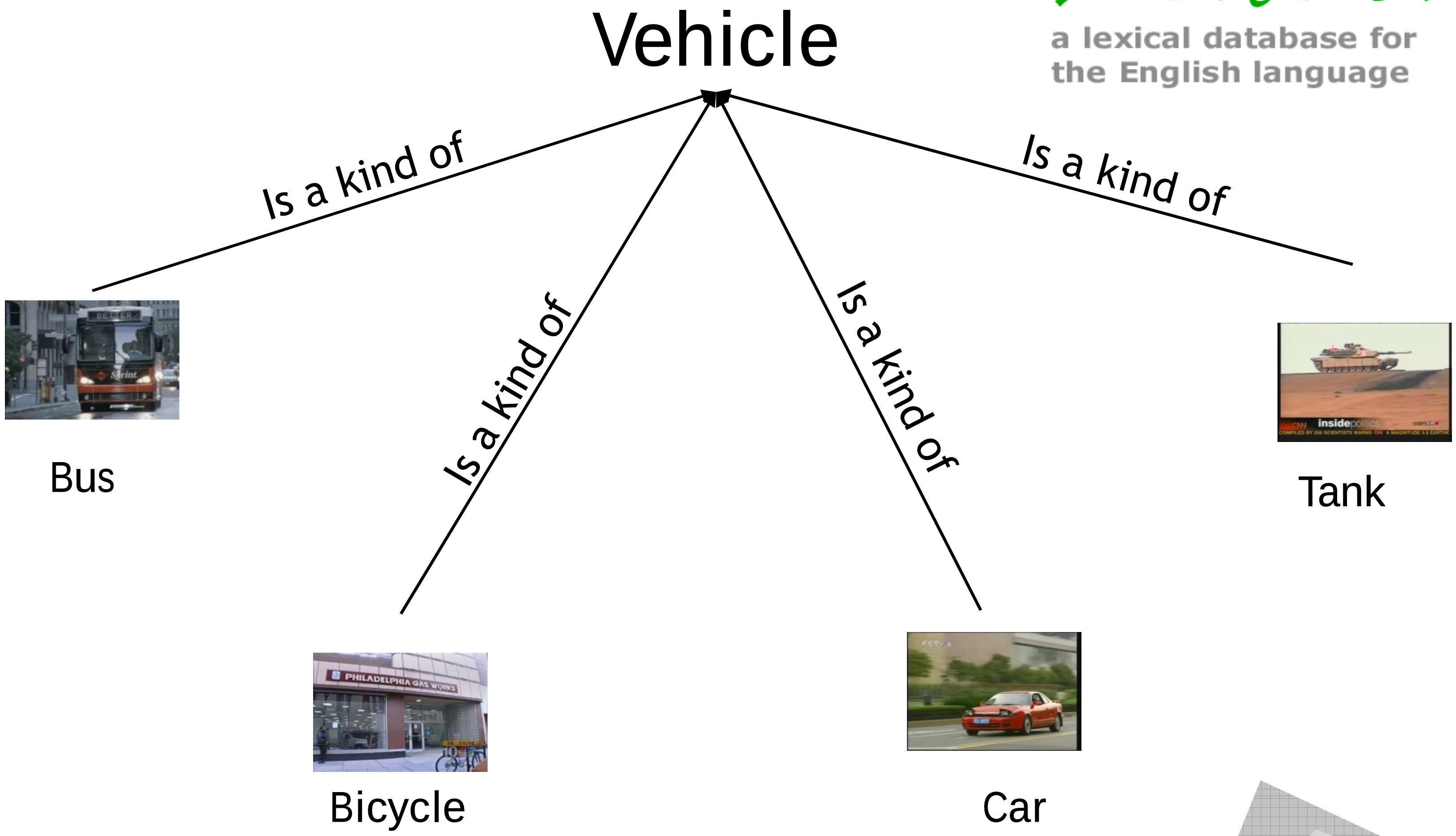
- Index concept descriptions
  - ✓ Represent as term vector
  - ✓ Only 363, so rather small collection
- Need to increase recall?
  - ✓ Porter stemming algorithm
  - ✓ Character *n*-gramming, here sequences of 4 characters
- We use the vector space model to match queries to descriptions
  - ✓ Pick detector that maximizes query/document similarity
- Turns out that perfect match yields best performance



WordNet relationships

- What
- Why
- How
- Why not?
- Conclusion

WordNet
a lexical database for the English language



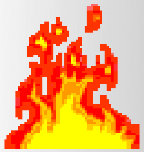
Ontology querying

- What
- Why
- How
- Why not?
- Conclusion

"Find a report from the desert showing a house or car on fire."

1. Identify objects in WordNet


car


fire


desert


house

Ontology querying

- What
- Why
- How
- Why not?
- Conclusion

"Find a report from the desert showing a house or car on fire."

2. Identify related concept detectors

car

fire

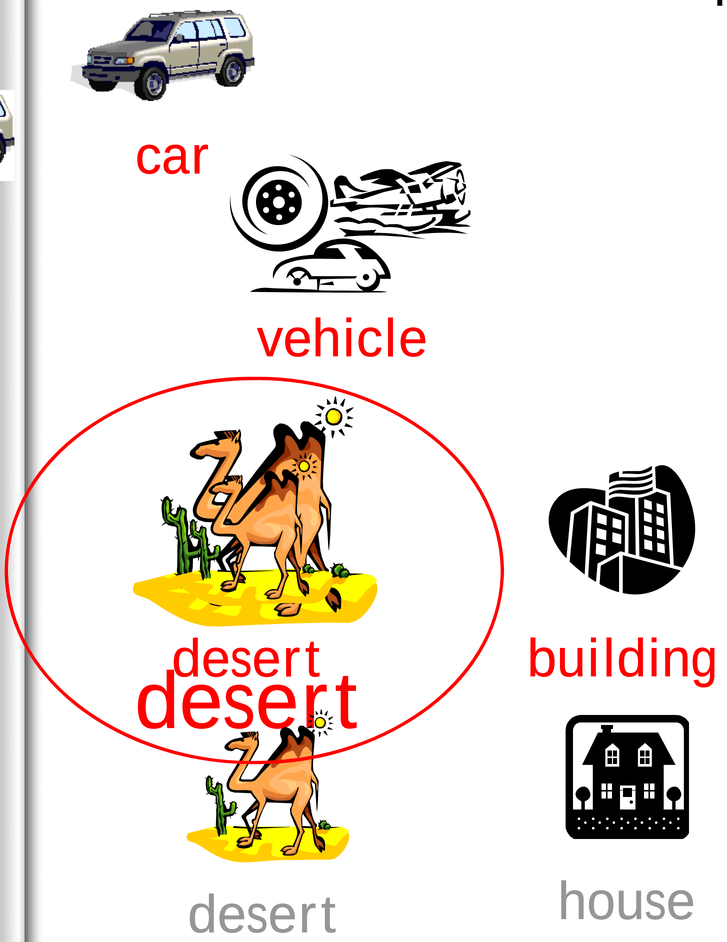
desert

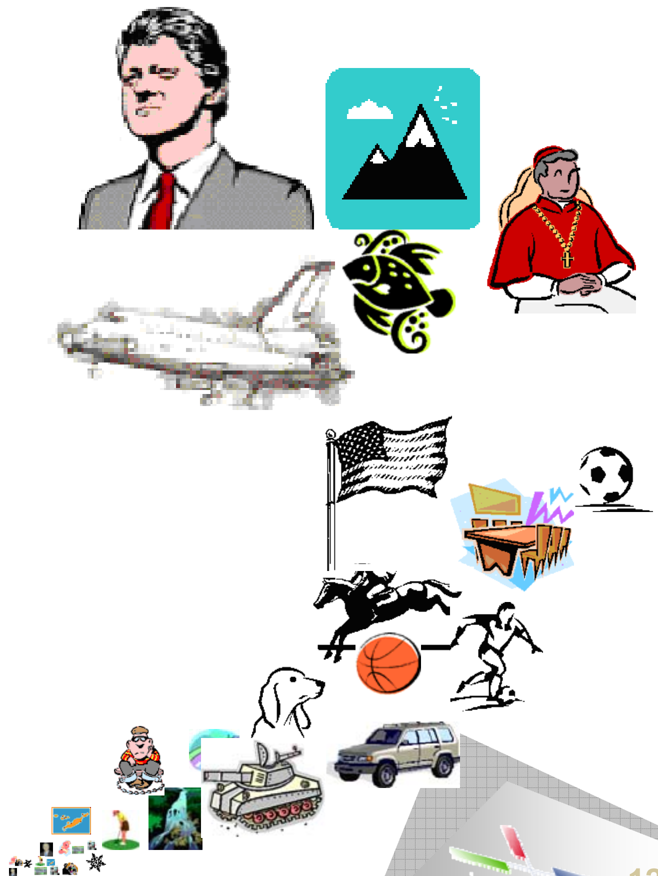
house

Ontology querying

“Find a report from the desert showing a house or car on fire.”

3. Find most similar and specific detector using Resnik’s measure

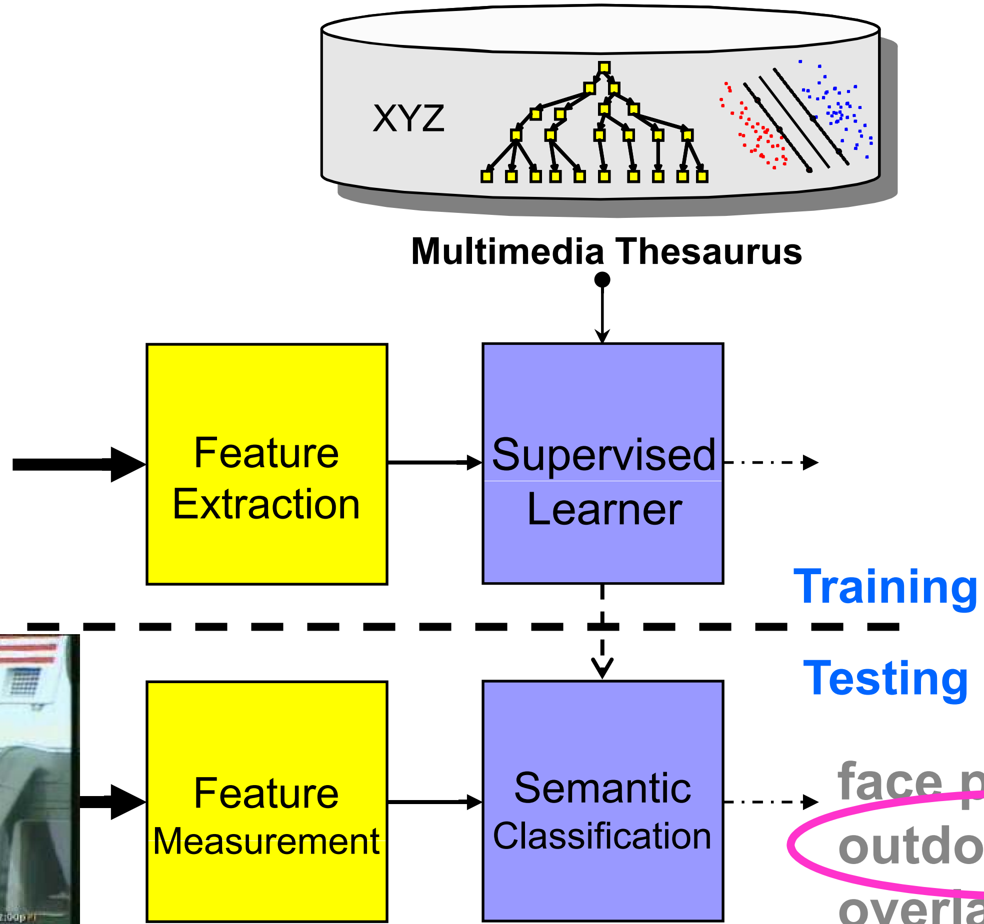




References:
Rasiwasia et al, TMM 9(5) 2007

Semantic visual querying

- What
- Why
- How
- Why not?
- Conclusion



Query Image

face p=0.97
outdoor p=0.98
overlayed text p=0.92
helicopter p=0.43
people p=0.96
...

Semantic visual querying

- Need to account for a priori concept occurrence
  - ✓ Prevent that robust/frequent concepts are selected
- Prioritize less frequent, but more discriminative concept detectors
  - ✓ Normalize posterior probability by concept occurrence
  - ✓ Estimate concept occurrence from manual annotations

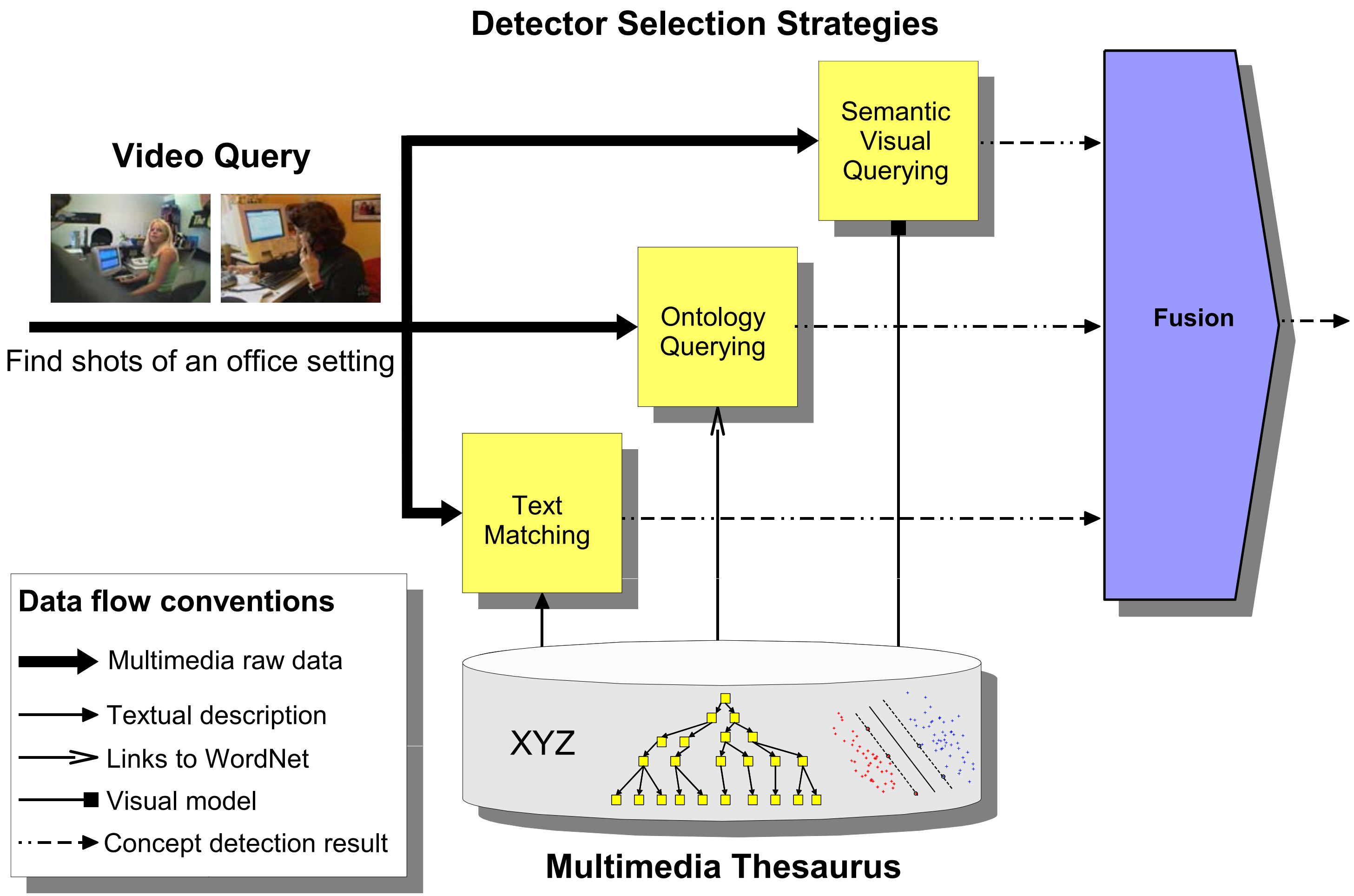
Query Image

	Posterior / Occurrence
face	p=0.97 / p=0.89
outdoor	p=0.98 / p=0.51
overlayed text	p=0.92 / p=0.75
helicopter	p=0.43 / p=0.09
people	p=0.96 / p=0.90
…	

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

Detector selection strategies

TRECVID 2005



Detector selection strategies

- What
- Why

		Best Possible		2a: Text Matching		2b: Ontology Querying		2c: Semantic Visual Querying
Search Topic	*Best Detector*	*AP*	*Selected Detector*	*AP%*	*Selected Detector*	*AP%*	*Selected Detector*	*AP%*
Two visible tennis players on the court	Athlete	0.6501	Tennis Game	89.7%	**Athlete**	**100.0%**	Tennis Game	89.7%
A goal being made in a soccer match	Stadium	0.3429	Soccer Game	31.7[illegible]	Soccer Game	31.7%	**Grass**	**51.0%**
Basketball players on the court	Indoor S[illegible]rts Venue	0.2801	Court	0[illegible]	Athlete	30.4%	**Basketball Game**	**81.5%**
A meeting with a large table and people	Furniture	0.1045	**Confer[illegible] Room**	[illegible]	Meeting	24.8%	Flag	1.0%
People with banners or signs	People Marching	0.1013	**Demo[illegible]r Protest**	[illegible]	Group	5.3%	Desert	0.4%
One or more military vehicles	Armored Vehicles	[illegible]892	**Tan[illegible]**	[illegible]	T[illegible]	**38.1%**	Charts	0.0%
Helicopter in flight	Helicopters	[illegible]	**Hel[illegible]**	[illegible]	[illegible]	**100.0%**	Helicopter Hovering	53.1%
A road with one or more cars	Car	[illegible]	[illegible]	[illegible]	[illegible]	**65.9%**	Helicopters	4.4%
An airplane taking off	Classroom	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**Helicopters**	**87.3%**
A tall building	Office [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Grass	0.2%
A ship or boat	Cloud	[illegible]	[illegible]	[illegible]	[illegible]	**[illegible]6.5%**	Cigar Boats	39.5%
George Bush entering or leaving vehicle	Rocket Propelled G[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**6.6%**	Helicopter Hovering	0.0%
Omar Karami	Chair	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]8%	**Yasser Arafat**	**3.5%**
Graphic map of Iraq, Baghdad marked	Graphical Map	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**Graphical Map**	**100.0%**
Condoleeza Rice	US National Fl[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0%	**Capitol**	**0.4%**
One or more palm trees	Weapons	0.02[illegible]	[illegible]	[illegible]	[illegible]	23.4%	**Fire Weapon**	**44.3%**
Something on fire with flames and smoke	Violence	0.0[illegible]	[illegible]	[illegible]	W[illegible]	41.4%	Soccer Game	18.9%
Mahmoud Abbas	Conference Room	0.[illegible]	[illegible]el	[illegible]	Ariel S[illegible]	0.5%	**Yasser Arafat**	**2.3%**
Hu Jintao	Iyad Allawi	[illegible]	Hu Jin[illegible]	[illegible]	George Bus[illegible]	2.4%	**Non-US National Flags**	**55.0%**
People shaking hands	Beards	0.0110	Handsh[illegible]	[illegible]%	Group	10.2%	**Yasser Arafat**	**18.0%**
Office setting	Computers	0.0095	**Compu[illegible]**	**1[illegible]%**	Office	90.4%	Emile Lahoud	1.9%
Iyad Allawi	Iyad Allawi	0.0095	**Iyad Al[illegible]**	**100.0%**	Ariel Sharon	46.6%	**Iyad Allawi**	**100.0%**
Tony Blair	Election Campaign Address	0.0067	Tony Bla[illegible]	0.0%	Tony Blair	0.0%	**George Bush jr**	**29.6%**
People entering or leaving a building	Muslims	0.0044	USA Government Building	6.4%	**Group**	**27.0%**	Reporters	8.5%
Mean		*0.0867*		*50.8%*		*56.0%*		*55.6%*
Number of highest scores		.		9		9		12

Detector selection varies
No best method
Why not fuse results?

Influence of detector selection combi

TRECVID 2005

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

Best | Text Matching | Ontology Querying | Visual Querying

Find shots of a tall building (with more than 5 floors above the ground)

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 one or more palm trees.

- What
- Why
- How
- Why not?
- Conclusion

Conclusions II

- Automatic video retrieval is a difficult problem
  - ✓ Many approaches exist, focusing on features and semantics
  - ✓ Results indicate that a multidisciplinary approach is most effective
  - ✓ Apart from low-level features and semantics, concept detectors should be part of solution
- Experiments with thesaurus of 363 concept detectors indicate:
  - ✓ 150 detectors sufficient to tackle 24 topics from TV2005
  - ✓ Selecting the right concept detector for a query depends on many facets: text, ontology, visual model
  - ✓ Combination of selection strategies potentially yields improved performance, but how to estimate the weights a priori?
- Conclusions cannot be definite as:
  - ✓ We only consider news domain, with very domain-specific detectors
  - ✓ We only consider 24 search topics, which is quite few
- Results do suggest promising new lines of research

VideOlympics

VideOlympics

Showcase Event

- What
- Why
- How
- Why not?
- Conclusion

- Benchmark performance cannot be sole criterion
  - ✓ Experience of searcher counts
  - ✓ Ease of use counts
  - ✓ …
- Demo session at TRECVID workshop shows a lot more than AP numbers
  - ✓ Individual demo's find their way to regular demo sessions
  - ✓ But they are never showed together again
- VideOlympics fills this lacuna
  - ✓ 'Live' interactive search task
  - ✓ Simultaneous exposure of video retrieval systems

What is it?

- A showcase that goes beyond the regular demo session
  - ✓ Fun to do for the participants
  - ✓ Fun to watch for the audience

- What should it be?
  - ✓ TRECVID participants simultaneously doing an interactive search task

Constraints

- What
- Why
- How
- Why not?
- Conclusion

- All systems should work on a laptop
  - ✓ No Internet connection
- Evaluation should be on the spot
- Minimal impact on existing systems
  - ✓ Software: simple submit mechanism
  - ✓ Data: TRECVID 2005/2006 only
- System performance not the deciding factor

Participants

Showcase Event

Columbia University in the City of New York
CuVid
digital video | multimedia laboratory

informedia
digital video understanding
Carnegie Mellon

Research Organization of Information and Systems
国立情報学研究所
National Institute of Informatics

FX PAL
FX Palo Alto Laboratory Inc.

Tsinghua University 1911
Intelligent Multimedia Group

IBM

MediaMill

VISIONGO
CO-DEVELOPED BY NUS&ICT

Setup

- What
- Why
- How
- Why not?
- Conclusion

case Event

One display

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

....

Was it fun?

VideOlympics

Showcase Event

- What
- Why
- How
- Why not?
- Conclusion



Golden Retriever Awards



http://www.flickr.com/groups/civr2007/pool/tags/videolympics/

Concept detection challenges

With Alex Hauptmann

- 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, …

- We believe you will have a hard time without solid knowledge of machine learning

- What
- Why
- How
- Why not?
- Conclusion

Concept retrieval challenges

With Alex Hauptmann

- 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?
- Do not assume a text query will give result
  - ✓ Consider home video domain for example
- How to balance semantic coverage and anticipated performance of detectors for a specific query?
- Have fun

- What
- Why
- How
- Why not?
- Conclusion

Acknowledgements

- University of Amsterdam
  - Richard van Balen, Jan van Gemert, Jan-Mark Geusebroek, Theo Gevers, Bouke Huurnink, Dennis Koelma, Giang P. Nguyen, Maarten de Rijke, Ork de Rooij, Koen van de Sande, Arnold Smeulders, Cor Veenman, Marcel Worring
- Free University Amsterdam
  - Laura Hollink, Guus Schreiber, Frank Seinstra
- Carnegie Mellon University
  - Alex Hauptmann
- IBM Research
  - Rong Yan

Further information

- What
- Why
- How
- Why not?
- Conclusion

➤ Including the sheets of the tutorial

www.MediaMill.nl