



Semantic Indexing & Retrieval of Video
Marcel Worring & Cees Snoek



with contributions by:
many

Intelligent Systems Lab Amsterdam,
University of Amsterdam, The Netherlands



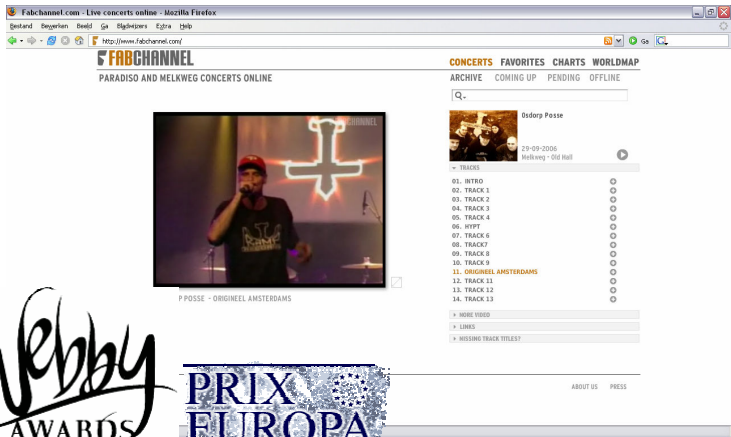
1. Amateur webcasts



- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval



- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

II. Professional webcasts



- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

III: Sport broadcasts



IV: News broadcasts

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

V: Science

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

VI: Homevideo

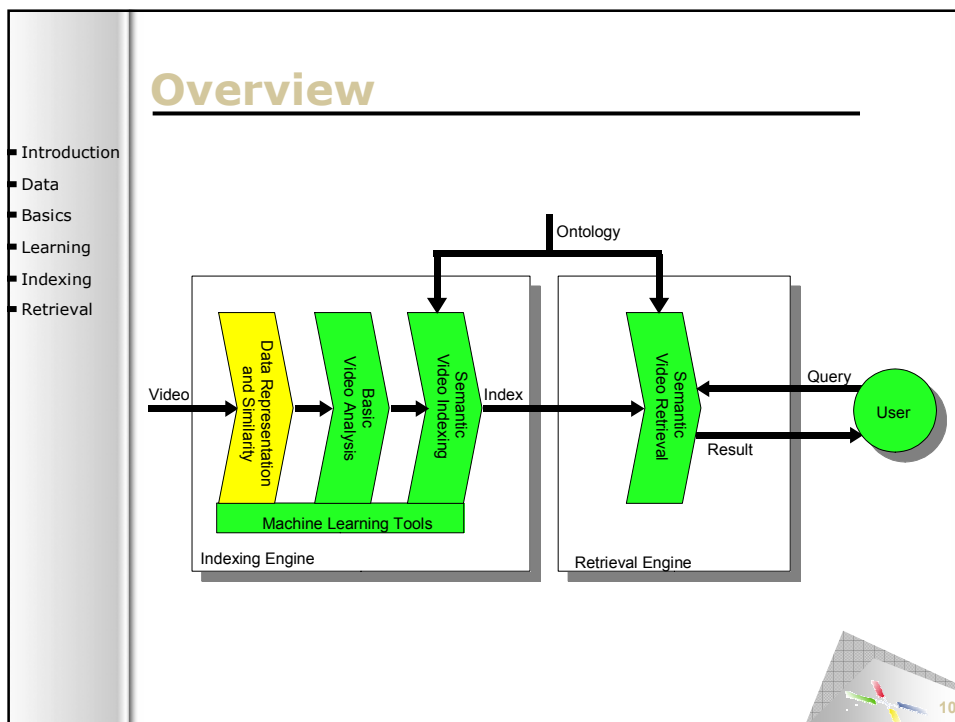
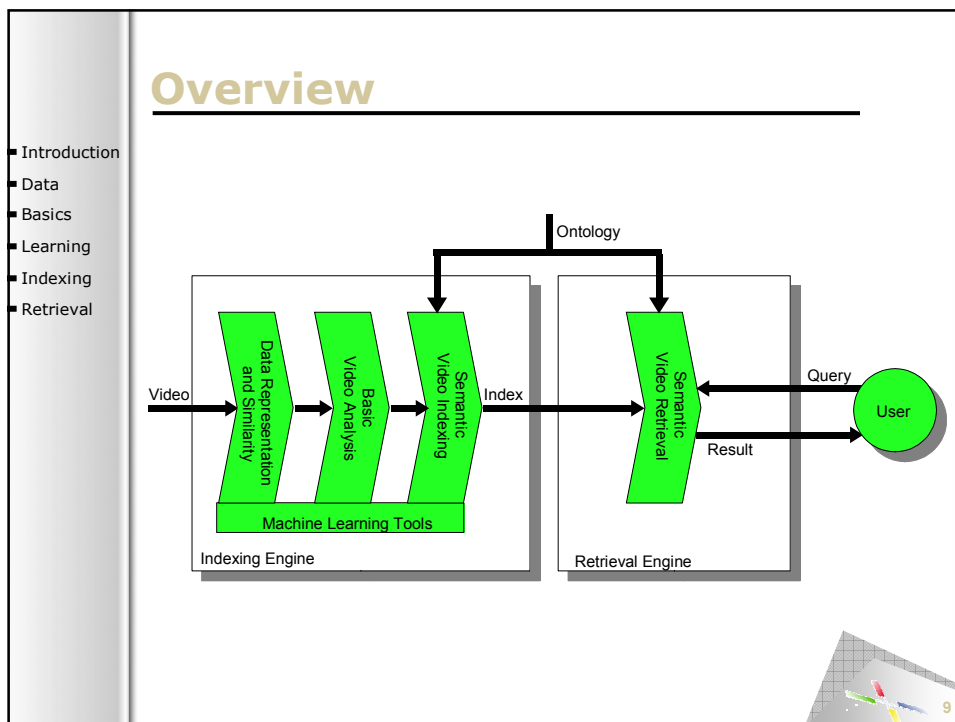
- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

7

VII. Surveillance


- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

8



- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Different data & needs



- Broadcasting
- E-Business
- Education
- Security
- Consumer Electronics

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

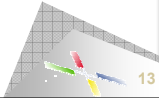
Data categories

- **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
 - ✓ Recorded 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.

Introduction
Data
Basics
Learning
Indexing
Retrieval

Factors of influence

- **Quality of the data**
 - ✓ What's the resolution and signal-to-noise-ratio of the data
- **Application control**
 - ✓ How much control does one have on the circumstances under which the data is recorded
- **Purpose**
 - ✓ The reason for which the video is being made being entertainment, information, communication, or data analysis

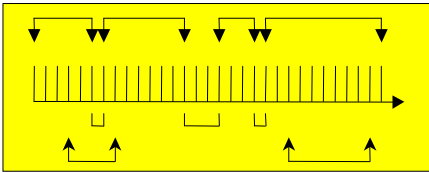


13

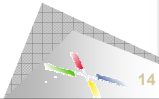
Introduction
Data
Basics
Learning
Indexing
Retrieval

The layout of the data

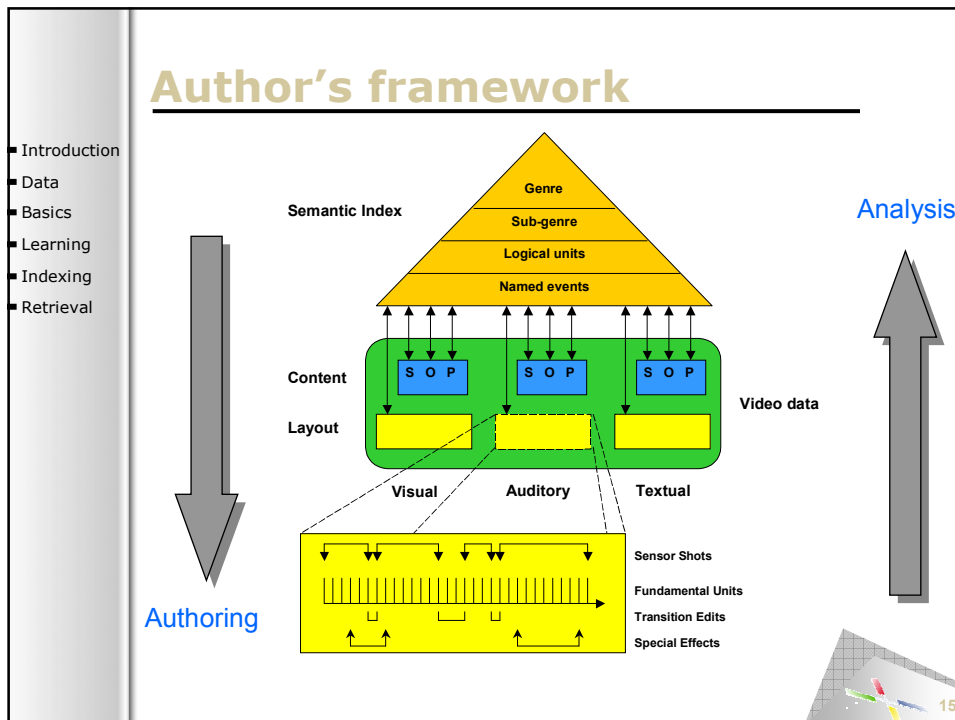
- **Fundamental units: general definition**
 - ✓ The elements in the video which are temporally ordered, but which do not have a temporal dimension themselves
- **Thus for the different modalities**
 - ✓ Visual: a frame
 - ✓ Auditory: a sample
 - ✓ Textual: a character
- **Characteristics**
 - ✓ Units usually do not have a meaning in isolation, it is their temporal aggregation that is important



Sensor Shots
Fundamental Units
Transition Edits
Special Effects



14



Non-produced video

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

- **Framework**
 - ✓ Remains valid, but not all elements are necessarily present e.g. security video does not have a layout
- **Analysis**
 - ✓ The analysis as presented considers recorded data and is hence very comprehensive, other domains are basically easier, but have often lower quality

16

Different low-level representations

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Regularity
 Coarseness
 Directionality

count Histogram

color

17

Color histogram

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

380 pixels

640 pixels

Total 243200 pixels

count Histogram

color

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

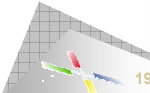
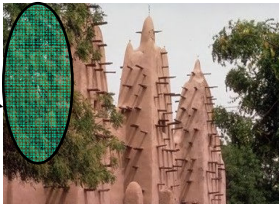
18

Introduction
Data
Basics
Learning
Indexing
Retrieval

Color coherence vectors

- Color histograms
 - ✓ Distribution of colors, no spatial information
- Color coherence vectors
 - ✓ Add an extra dimension counting how many pixels in the neighborhood of the pixel have the same color, so the CCV states how many pixels of color x with y neighbors with the same color are present in the image

Large "green" blob

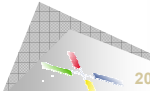
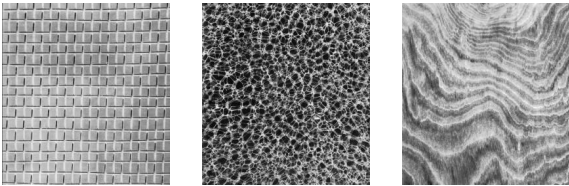


19

Introduction
Data
Basics
Learning
Indexing
Retrieval

Texture

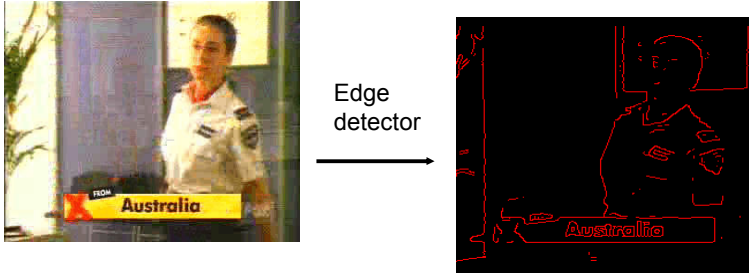
- Color of a pixel
 - ✓ Defined for the one pixel only
- Texture of a region
 - ✓ The pattern induced by the different colors/intensities in a neighborhood
 - ✓ Measures
 - ❖ Coarseness: are the basis elements small or big?
 - ❖ Directionality: can we observe a line like pattern?
 - ❖ Regularity: is there some basis form that is repeated?



20

Introduction
Data
Basics
Learning
Indexing
Retrieval

Edges



Edge detector

Summary of the data selecting positions in the image where there is a strong change in color (or texture).

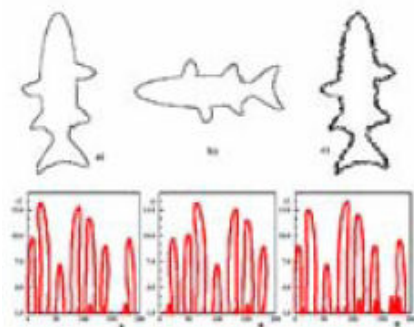
21

The slide illustrates the edge detection process. On the left, a photograph of a woman in a white shirt with a yellow 'Australia' banner is shown. An arrow labeled 'Edge detector' points to the right, where the same image is shown with its edges highlighted in red on a black background. The edges of the woman's face, hair, and the banner are clearly visible.

Introduction
Data
Basics
Learning
Indexing
Retrieval

Shape

- **The curvature scale space**
 - ✓ A commonly used descriptor to capture both the global and more detailed shape of an object
- **Most shape descriptors**
 -
 - ✓ Only work if you have a nicely delineated object, either manually extracted or objects on uniform background, and seen from the same viewpoint



22

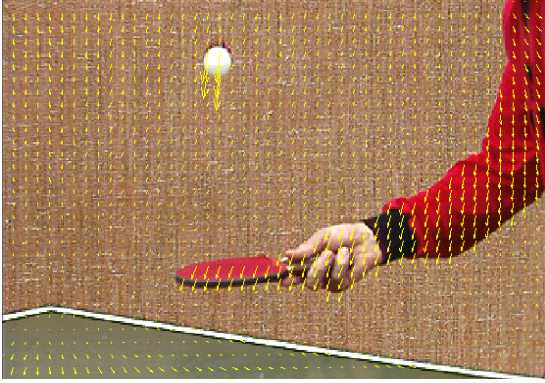
The slide discusses shape descriptors. It features three fish outlines: a side view of a fish, a top-down view of a fish, and a side view of a different fish. Below each outline is a corresponding 'curvature scale space' plot, which is a red line graph showing the curvature of the object's boundary at various points. The plots show how the curvature changes as you move along the perimeter of the fish.

Introduction
Data
Basics
Learning
Indexing
Retrieval

Motion

➤ The optic flow field

- ✓ Indicates where pixels are moving to in the next frame can e.g. be summarized by considering
 - ❖ the dominant motion or the average speed

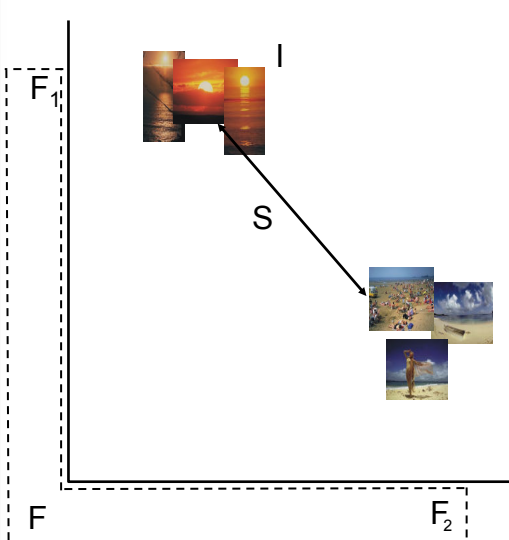


The image shows a hand holding a red object against a textured background. A grid of yellow arrows represents the optic flow field, indicating the direction and speed of pixel movement from one frame to the next. The arrows generally point downwards and to the right, consistent with the hand's position and the background's texture.

23

Introduction
Data
Basics
Learning
Indexing
Retrieval

From descriptors to similarity



The diagram illustrates the process of mapping an image I to a feature space F_1 and then to a similarity space F_2 . An arrow labeled S points from the image I to a set of smaller images in the F_2 space, representing similarity-based retrieval.

Many possible functions


- Color based
- Shape based
- Texture based
- Layout based
- Semantics based

24

The goal: semantic video indexing

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

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

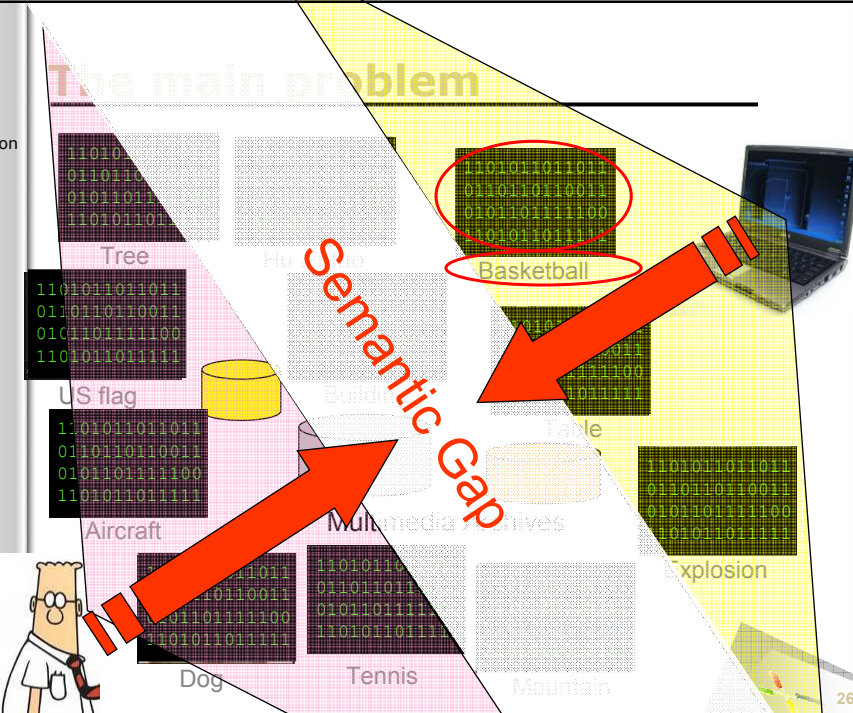


Airplane

25

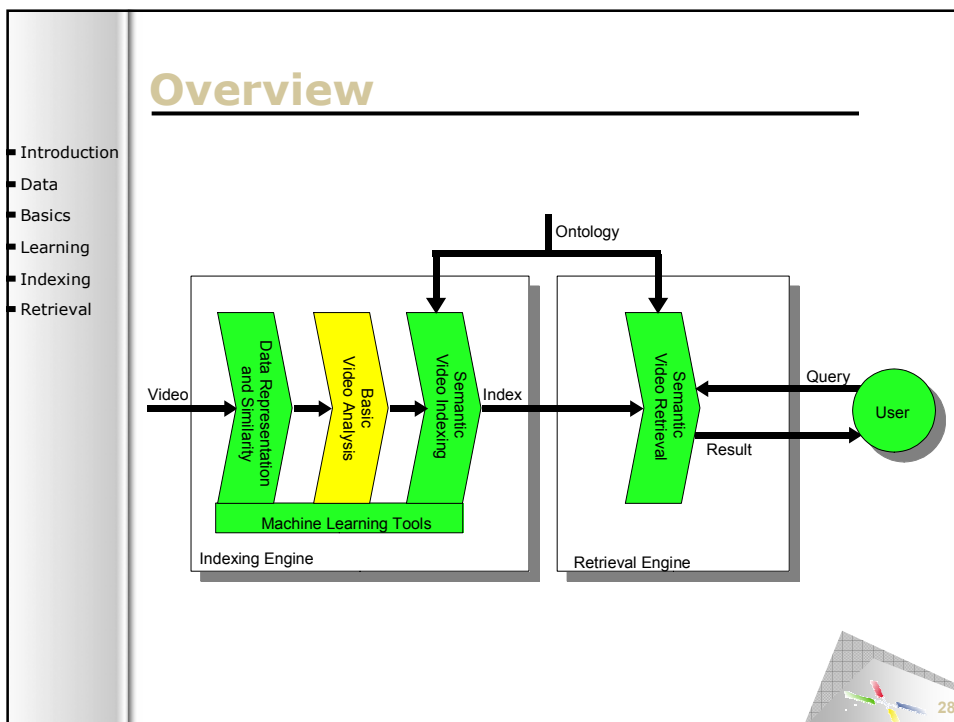
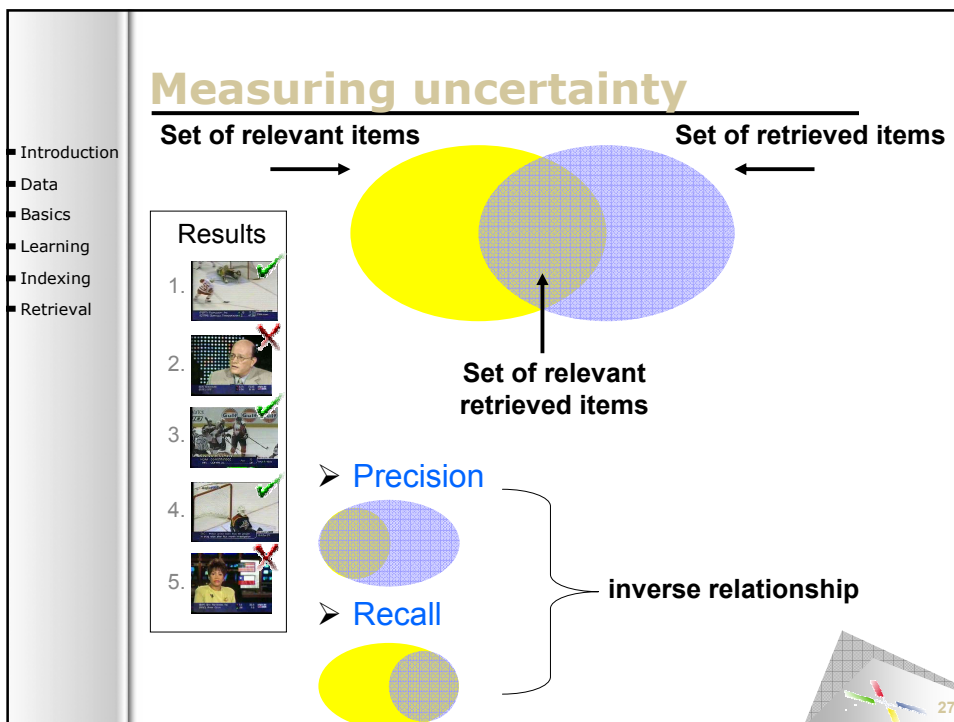
The main problem

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval



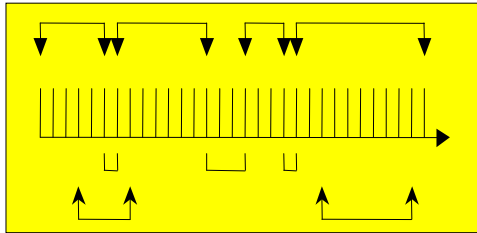
Semantic Gap

26

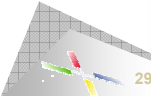


The layout

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval



The diagram shows a yellow rectangular area representing a video layout. At the top, there are several downward-pointing arrows of varying lengths, labeled 'Sensor Shots'. Below these is a horizontal bar divided into many small segments, labeled 'Fundamental Units'. Underneath the bar, there are several upward-pointing arrows of varying lengths, labeled 'Transition Edits'. At the bottom, there are several upward-pointing arrows of varying lengths, labeled 'Special Effects'.

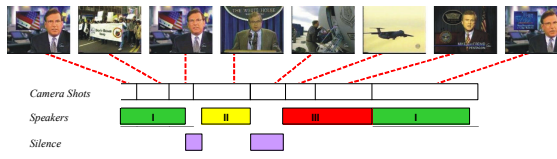


29

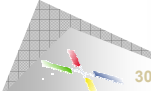
Overview: layout reconstruction

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

- Detection of shots and transitions
- Visual
 - ✓ Detection of shot cuts and gradual effects
- Auditory
 - ✓ Detection of silence and transition points



The diagram illustrates layout reconstruction. At the top, a row of eight small video frames shows different scenes. Below them is a horizontal timeline. Red dashed lines connect the frames to the timeline. Under the timeline, there are three colored bars: a green bar labeled 'I', a yellow bar labeled 'II', and a red bar labeled 'III'. Below these bars, there are two purple squares labeled 'Silence'.



30

Shot detection

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval




A shot is the result of a continuous camera operation and which doesn't contain an edit


31


Histogram based analysis

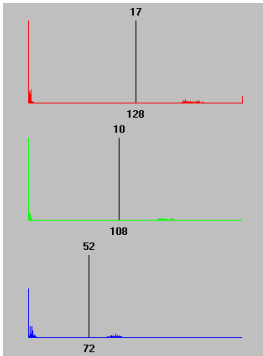
- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Frame t

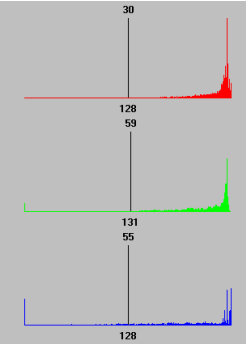


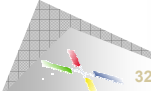
Frame t+1





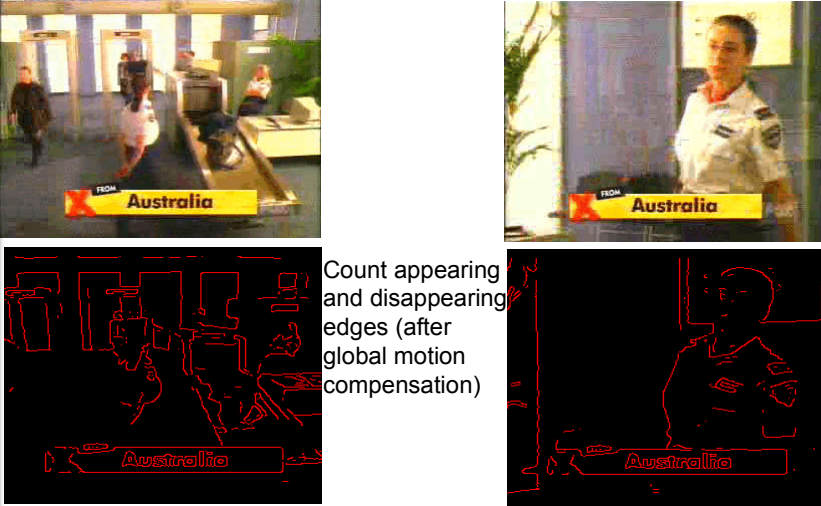
Difference of histogram



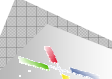

32

Shot detection: object based

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

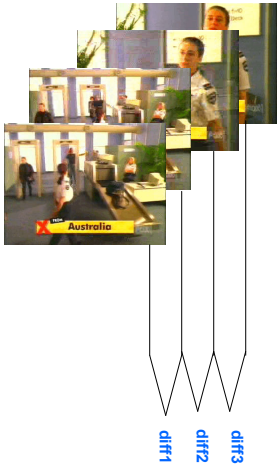


Count appearing and disappearing edges (after global motion compensation)

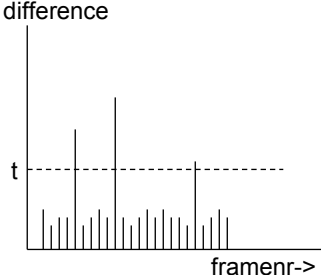

33

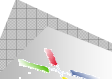
Shot detection

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval



difference




34

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Cut detection requirements

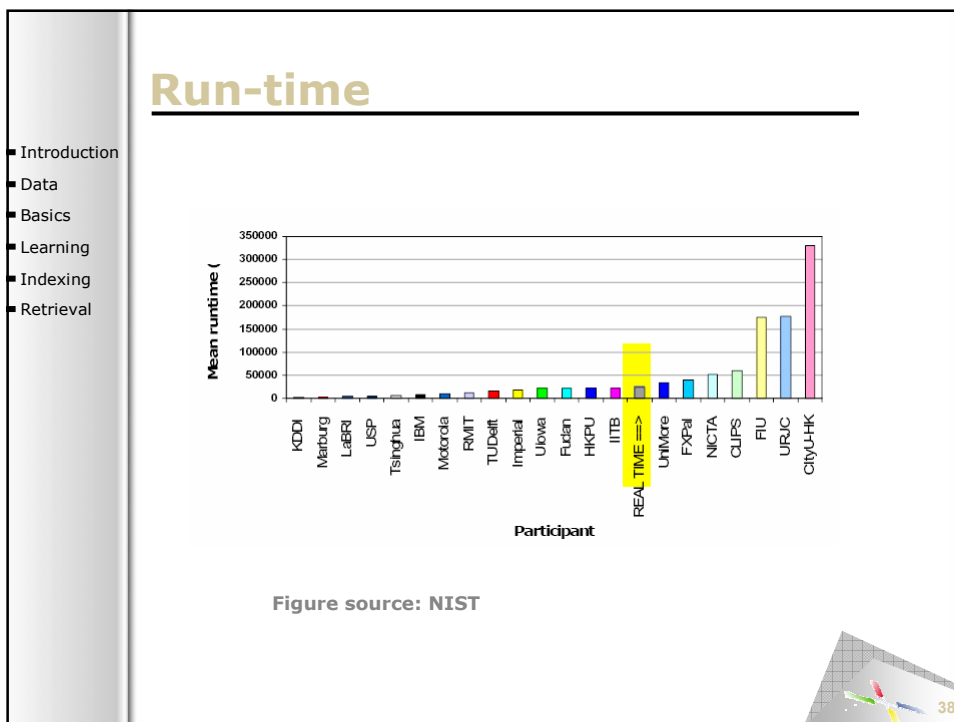
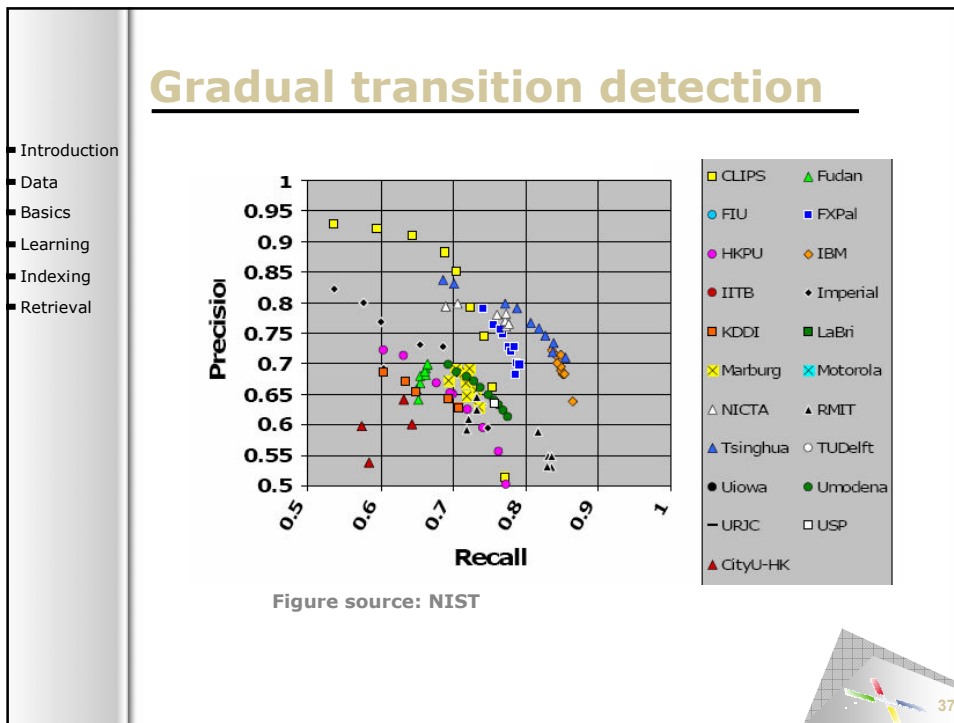
- **robust to camera movement**
 - ✓ gradual changes of camera position or focus should have no or little influence on the difference metric
- **robust to object movement**
 - ✓ objects moving in the scene should marginally affect metric
- **robust to edit effects**
 - ✓ if we only want to detect hard cuts, dissolves and other edit effects should not affect the result
- **robust to changes in lighting conditions**
 - ✓ changes in lighting like flashes, different viewing angle, differently colored lights etc. can have a large effect on the colors in the image, should be of minor influence
- **efficient**
 - ✓ should be computed very fast as the number of frames to process is very large and method should be preferably real-time

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Cut detection performance

CLIPS	Fudan
FIU	FXPal
HKPU	IBM
ITB	Imperial
KDDI	LaBri
Marburg	Motorola
NICTA	RMIT
Tsinghua	TU Delft
Uowa	Uomodena
URJC	USP
CityUHK	

Figure source: NIST

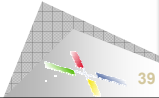


Introduction
Data
Basics
Learning
Indexing
Retrieval

Layout reconstruction: audio

➤ Abrupt cuts

- ✓ detection of silence
 - ❖ simple method, just consider the average energy in a window
 - ❖ clearly silence segments have low energy
- ✓ more advanced: detection of positions where there is a transition from one category of sound to the other
 - ❖ detect points where there is a clear increase in energy
 - ❖ detect points where there is a clear decrease in energy
 - ❖ both detected by moving a window over the signal and by comparing the left and right side of the window
 - ❖ merge adjacent breaks of the same type



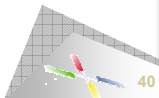
39

Introduction
Data
Basics
Learning
Indexing
Retrieval

Layout reconstruction: audio

➤ Music detection

- ✓ goal: distinguish music from speech, silence, and environmental sounds
- ✓ based on the following four features
 - ❖ the harmonicity
 - ❖ Local measures of frequency

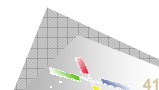




40

Introduction
Data
Basics
Learning
Indexing
Retrieval

VideoOCR

- **Caption: definition**
 - ✓ any text overlaid on the video
- **Caption detection assumptions**
 - ✓ uniform colour and brightness
 - ✓ clear character edge
 - ✓ disjoint characters
 - ✓ stationary and horizontally aligned text

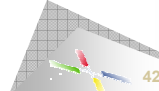



41

Introduction
Data
Basics
Learning
Indexing
Retrieval

Caption detection

- For each frame compute the number of frames with approximately the same colour in a part of the next frame
- Mark the frames in which this number falls in a certain range
- search for a sequence of sufficient length



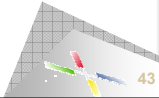
42

Introduction
Data
Basics
Learning
Indexing
Retrieval

Caption recognition

- Recognition of detected text
 - ✓ VideoOCR similar to text documents
 - ✓ (OCR = Optical Character Recognition)

A	→	A
U	→	u
S	→	s
t	→	I,t,l ??

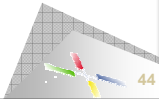



43

Introduction
Data
Basics
Learning
Indexing
Retrieval

Caption recognition

- Characteristics of video text are quite different from regular printed text hence VideoOCR is not OCR
 - ✓ Different fonts used
 - ❖ System has to be trained on the specific fonts
 - ✓ Confusion not a consequence of typing
 - ❖ (e.g. close on the keyboard)
 - ✓ Often isolated strings, many statistics
 - ❖ Hence context has to be used in a different way



44

Speech recognition

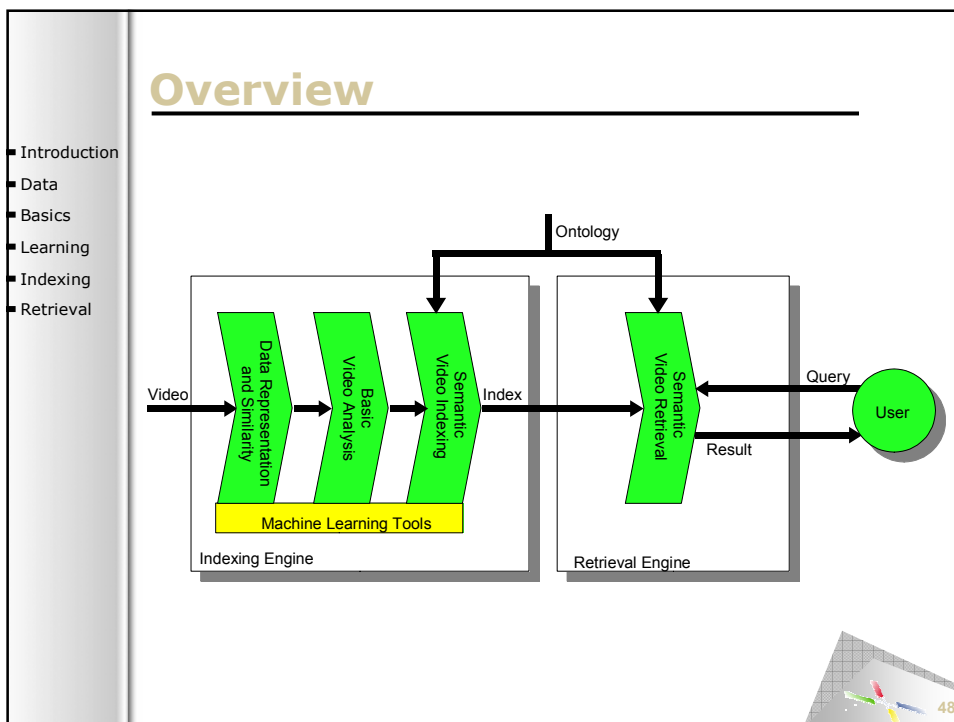
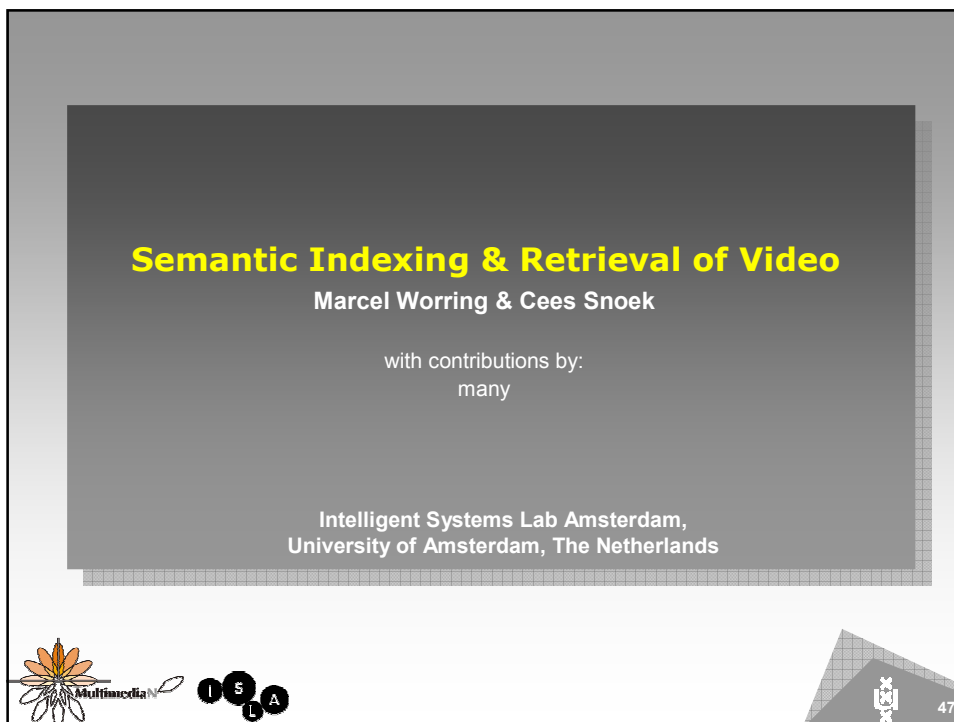
- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

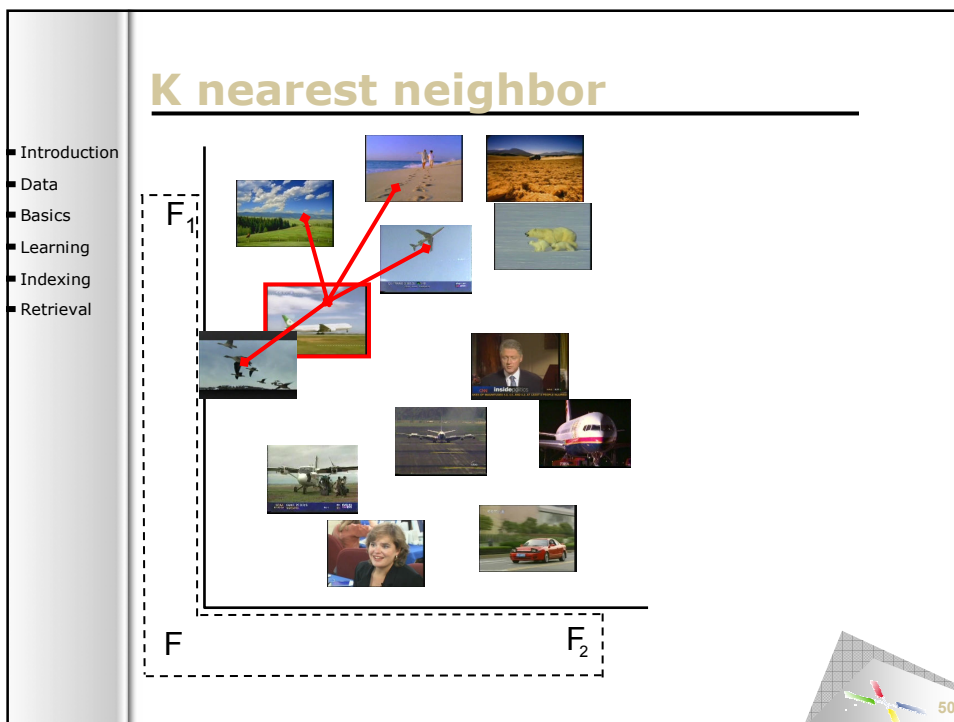
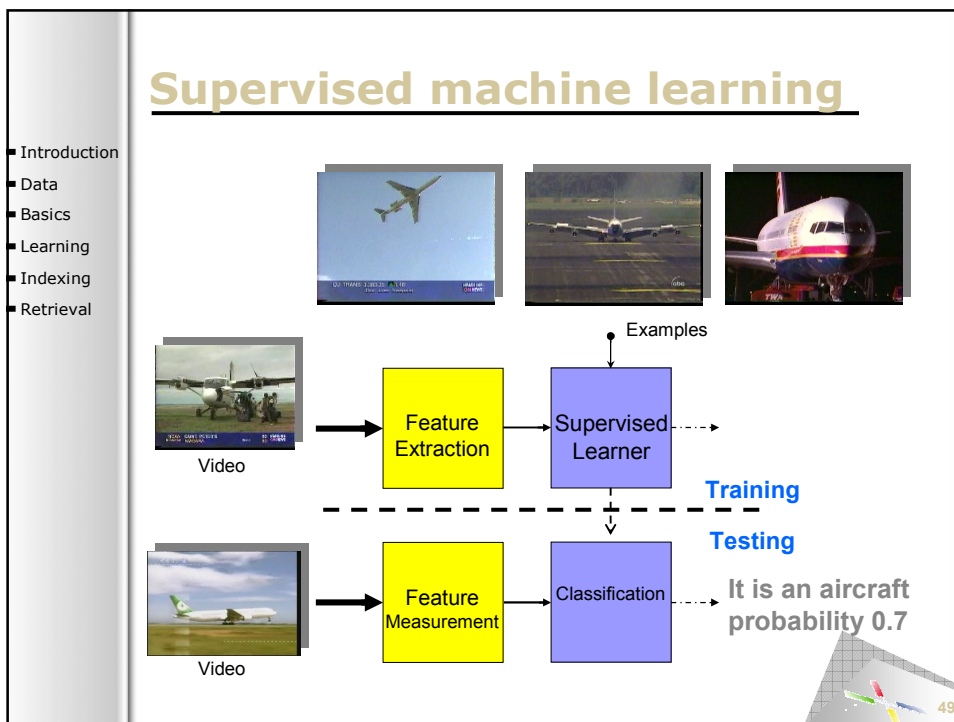
- **The state-of-the-art**
 - ✓ Works well on signals with high signal-to-noise ratio with a more or less fixed vocabulary with a limited number of known speakers
 - ✓ American English a lot better than Chinese, Arabic, or Dutch; mostly because of the amount of training data available
- **Note**
 - ✓ This is a complete research field of its own

45

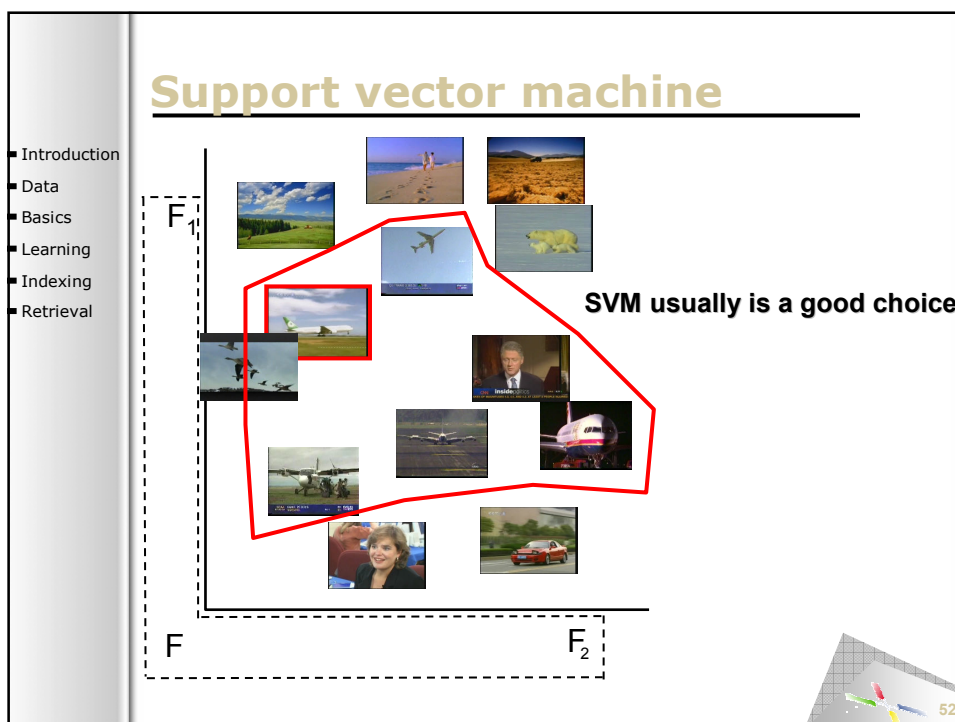
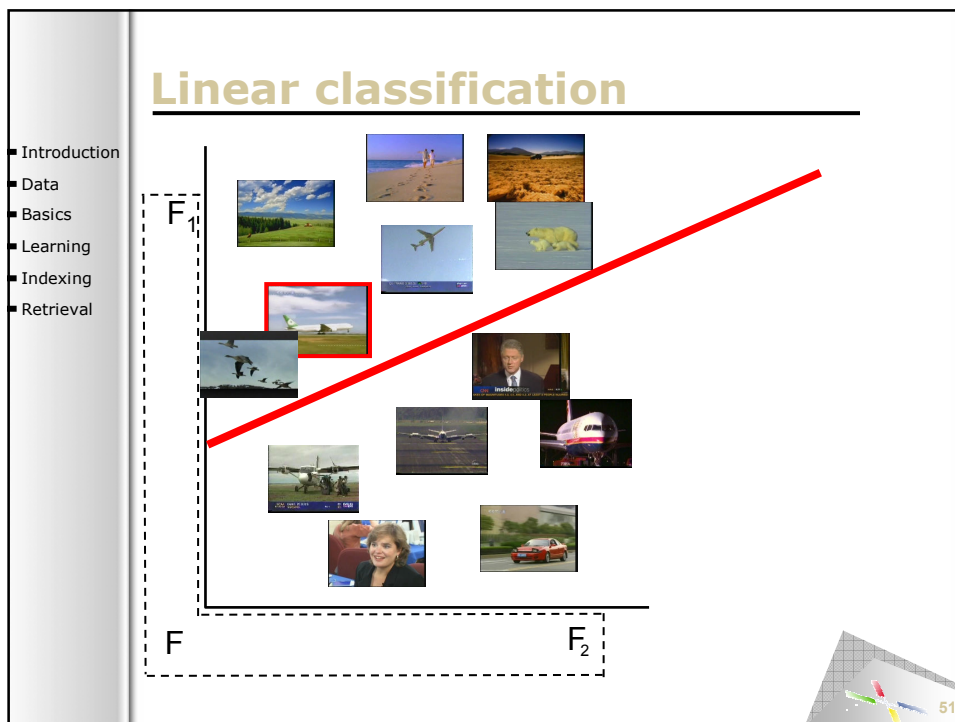
Author's framework

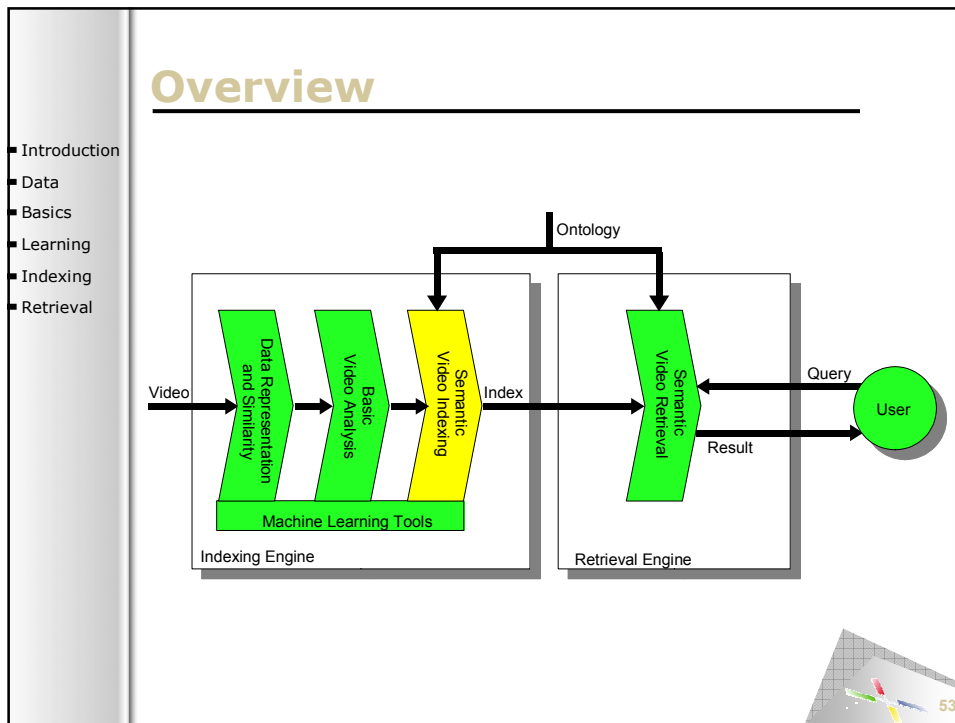
46





- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval





Semantic indexing

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

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

JOCELYN DEFORE

A face detector for frontal faces

↓ 3 years later






A face detector for non-frontal faces






One (or more) PhD for every new_concept





54

So how about these?

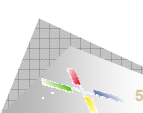
- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

And the > 1000 others

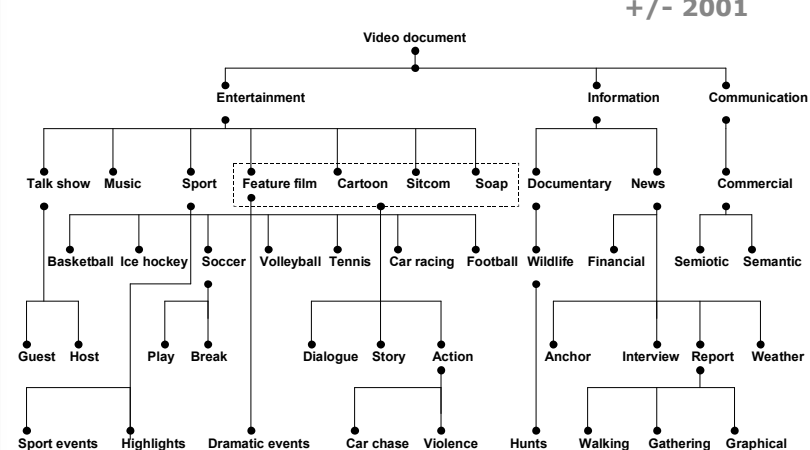


55

Semantic index overview

+/- 2001

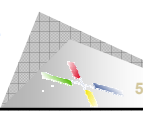
- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval



```

graph TD
    VD[Video document] --> Ent[Entertainment]
    VD --> Inf[Information]
    VD --> Com[Communication]
    
    Ent --> TS[Talk show]
    Ent --> Mus[Music]
    Ent --> Sport
    Ent --> FF[Feature film]
    Ent --> Cart[Cartoon]
    Ent --> Sit[Sitcom]
    Ent --> Soap
    
    Sport --> B[Basketball]
    Sport --> IH[Ice hockey]
    Sport --> Soc[Soccer]
    Sport --> Vol[Volleyball]
    Sport --> Ten[Tennis]
    Sport --> CR[Car racing]
    Sport --> Foot[Football]
    
    FF --> G[Guest]
    FF --> H[Host]
    FF --> P[Play]
    FF --> Bk[Break]
    FF --> D[Dialogue]
    FF --> S[Story]
    FF --> A[Action]
    
    A --> SE[Sport events]
    A --> Hl[Highlights]
    A --> DE[Dramatic events]
    A --> CC[Car chase]
    A --> V[Violence]
    
    Inf --> Doc[Documentary]
    Inf --> News
    
    Doc --> W[Wildlife]
    Doc --> Fin[Financial]
    
    News --> An[Anchor]
    News --> Int[Interview]
    News --> R[Report]
    News --> We[Weather]
    
    Com --> C[Commercial]
    
    C --> Sem[Semiotic]
    C --> Sem2[Semantic]
    
    Sem --> Hn[Hunts]
    Sem --> Wk[Walking]
    Sem --> Gt[Gathering]
    Sem --> Gr[Graphical]
    
```

One PhD per detector requires too many students...

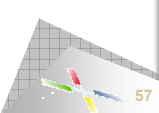


56

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

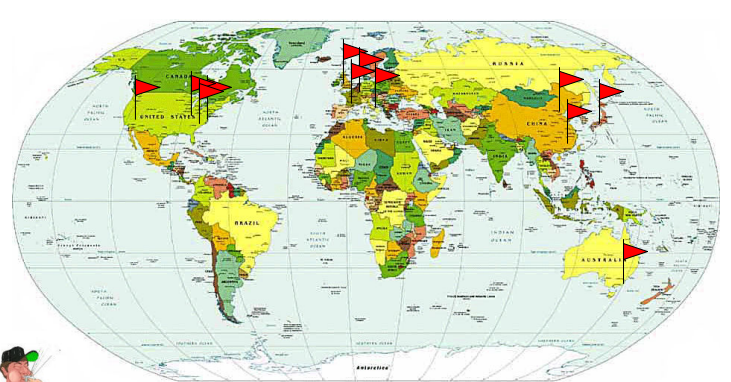
By 2001, state of affairs


Target	Note
Layout segmentation	cut detection solved for regular footage
Content detection	mostly unimodal analysis
Identify a concept class	<i>one Ph.D per detector</i> <ul style="list-style-type: none"> • face detector • car detector • boat detector • tree detector
Data sets used	<< 10 hours
Are we making progress?	<i>Impossible to measure...</i>


57

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

How about performance?



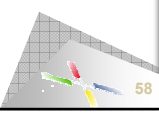


NIST

▶ Video analysis researchers

- ✓ Until 2001 everybody uses **specific** and **small** data sets
- ✓ Hard to compare methodologies

Since 2001 worldwide evaluation by NIST


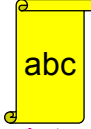
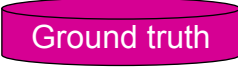

58

<http://www-nlpir.nist.gov/projects/trecvid/>

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**

59

Concept detection task


- **Given a video dataset and N semantic concept definitions:**
 - ✓ How well can you detect the concepts?

Growing Participation

TRECVID 2001 12 Participants 11 Hours of NIST video	TRECVID 2002 17 Participants 73 Hours of Video from Prelinger archives	TRECVID 2003 24 Participants 133 Hours of 1998 ABC, CNN news & C-SPAN	TRECVID 2004 38 Participants 173 Hours of 1998 ABC, CNN news & C-SPAN	TRECVID 2005 62 Participants 220 Hours of 2004 news from U.S., Arabic, Chinese sources, BBC stock shots
---	--	---	---	---

Growing Data Sets



...

Figure source: IBM

60

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

TRECVID evaluation measures

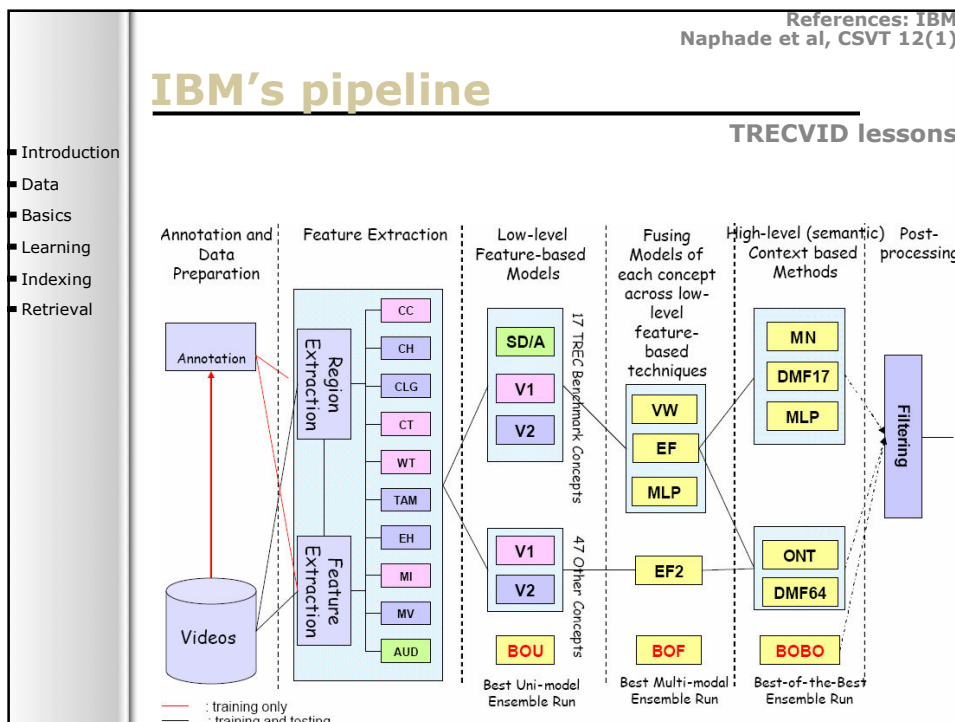
- **Classification procedure**
 - ✓ Training: many hours of (partly) annotated video
 - ✓ Testing: many hours of **unseen** video
- **Evaluation measure: Precision at N**
 - ✓ How many items correct in the first N results
- **Evaluation measure: Average Precision**
 - ✓ Combines precision and recall
 - ✓ Averages precision after every relevant shot
 - ✓ Top of the ranked list most important

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

Results

- 1.
- 2.
- 3.
- 4.
- 5.

61



References: IBM
Naphade et al, CSVT 12(1)

IBM's pipeline approach

TRECVID lessons

- 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,...

63

References: IBM
Naphade et al, CSVT 12(1)

IBM's pipeline approach

TRECVID lessons

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

64

References: IBM
Naphade et al, CSVT 12(1)

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

65

References: IBM

Concept detection task

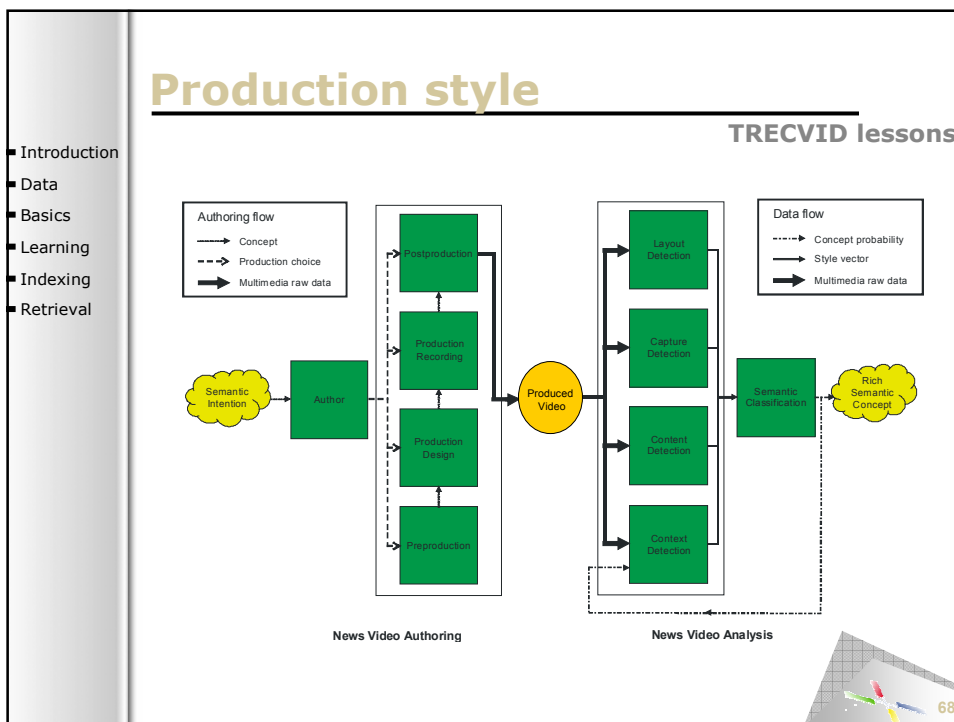
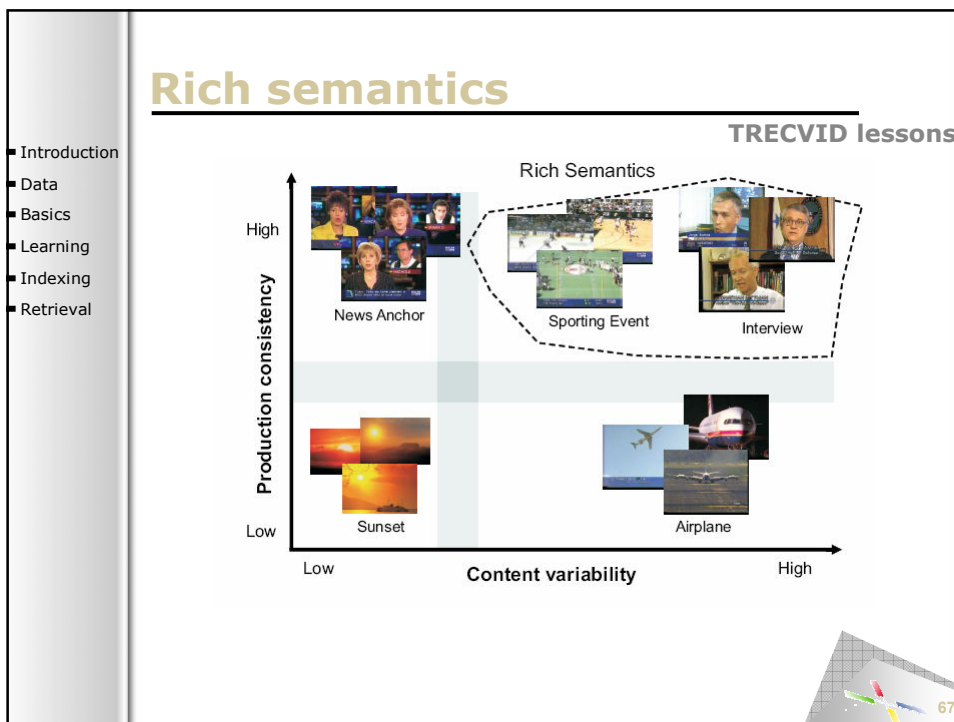
TRECVID 2003

Category	Best IBM	Best Non-IBM
Outdoors	0.2	0.15
NS_Face	0.15	0.1
People	0.25	0.15
Building	0.15	0.1
Road	0.15	0.1
Vegetation	0.15	0.1
Animal	0.35	0.25
Fence_Speech	0.2	0.15
Car/Truck/Bus	0.2	0.15
Aircraft	0.4	0.2
NS_Monochrome	0.3	0.2
NonSub_Selfie	0.1	0.6
Sporting	0.15	0.1
Weather	0.7	0.35
Zoom_In	0.85	0.85
Physical_Violence	0.1	0.6
Machine_Airflight	0.1	0.1
Mean	0.35	0.2

■ IBM (Generic)
■ Others (Specific)

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

66



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

Style detectors

TRECVID lessons

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Camera Shots

Speakers

Voice over

Frequent

Silence

69

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

Style detectors

TRECVID lessons

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

“Ann Compton
ABC news,
New York”

match?

Jemmy Curter

isName?

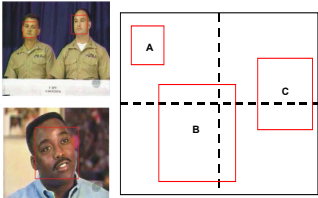
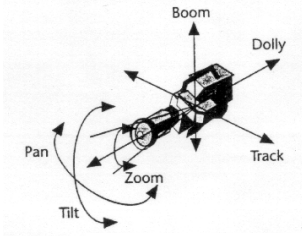

“Call now:
1800...”

70

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

Style detectors

TRECVID lessons

71

TRECVID 2003

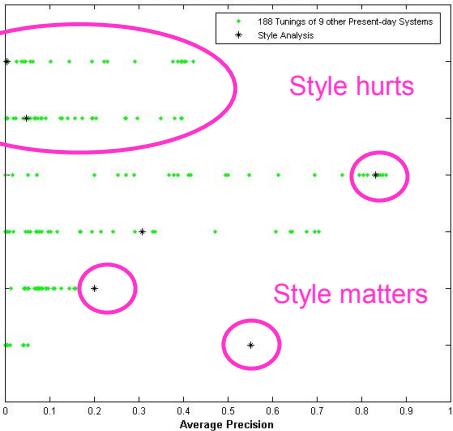
Concept detection task

Benchmark Comparison

- Others
- * Style

Style hurts

Semantic Concept



Average Precision

72

Key frame based analysis

TRECVID lessons

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

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?

73

A visual analysis scenario

TRECVID lessons

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

18 Setting Patches → Pixel Labeling → Vector → SVM → Concepts

Vector: 11% Greenery, 15% Concrete, ... 18% Sky

SVM: Labeled examples

Concepts: Outdoor?, Road?, Car?


74

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Frame combination functions


TRECVID lessons

Shot




Frame 1

$P(car) = 0.81$



Frame 2

$P(car) = 0.78$



Frame n

$P(car) = 0.77$

Shot-based combination: $1/n \sum P(car | frame)$

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

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

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Concept detection surplus value

TRECVID lessons

AP

weather news (0.827)

basket soccer (0.545)

stock quotes (0.472)

overlaid text (0.405)

people walking (0.283)

graphics (0.243)

cartoon (0.230)

physical violence (0.211)

people (0.188)

vegetation (0.170)

financial news anchor (0.077)

ice hockey (0.070)

Bill Clinton (0.039)

dog (0.032)

berlin (0.028)

antlers (0.027)

car (0.024)

airplane take off (0.023)

peel (0.022)

American football (0.018)

baseball (0.017)

host (0.007)

train (0.005)

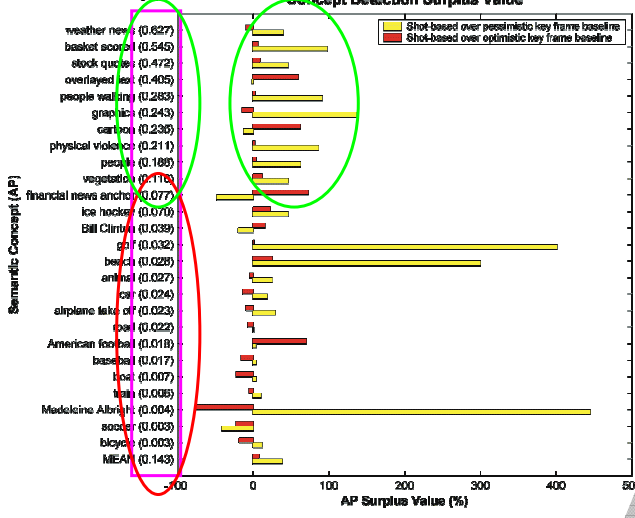
Medicine Albert (0.004)

sunset (0.003)

bicycle (0.003)

MEAN (0.143)

Concept Detection Surplus Value




Legend:
■ Shot-based over pessimistic key frame baseline
■ Shot-based over optimistic key frame baseline

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

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.






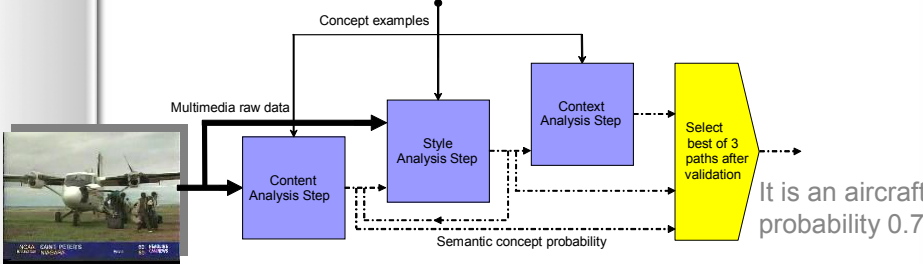
Integrated architecture principled on authoring metaphor

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

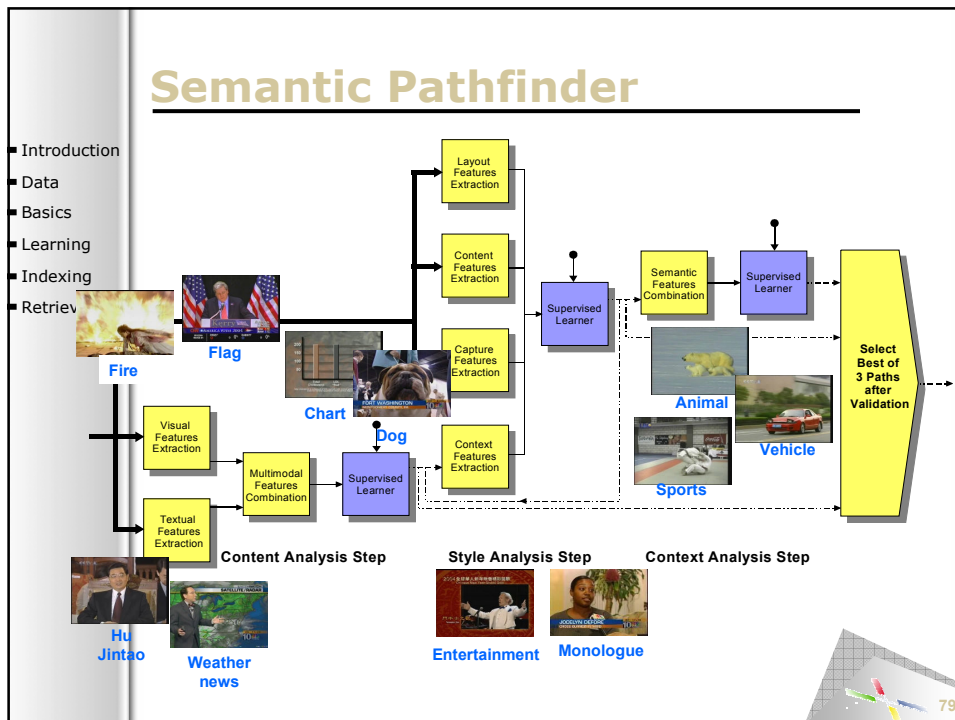
Semantic Pathfinder

following the authoring metaphor



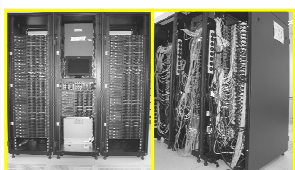
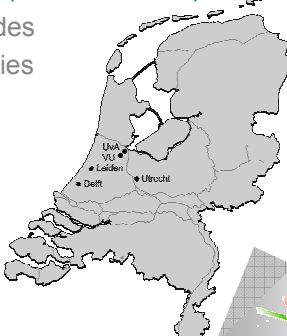
It is an aircraft probability 0.7



References:
Seinstra et al, IPDPS, 2005

How to analyze large video archives?

- **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

Annotated 32 concept lexicon

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

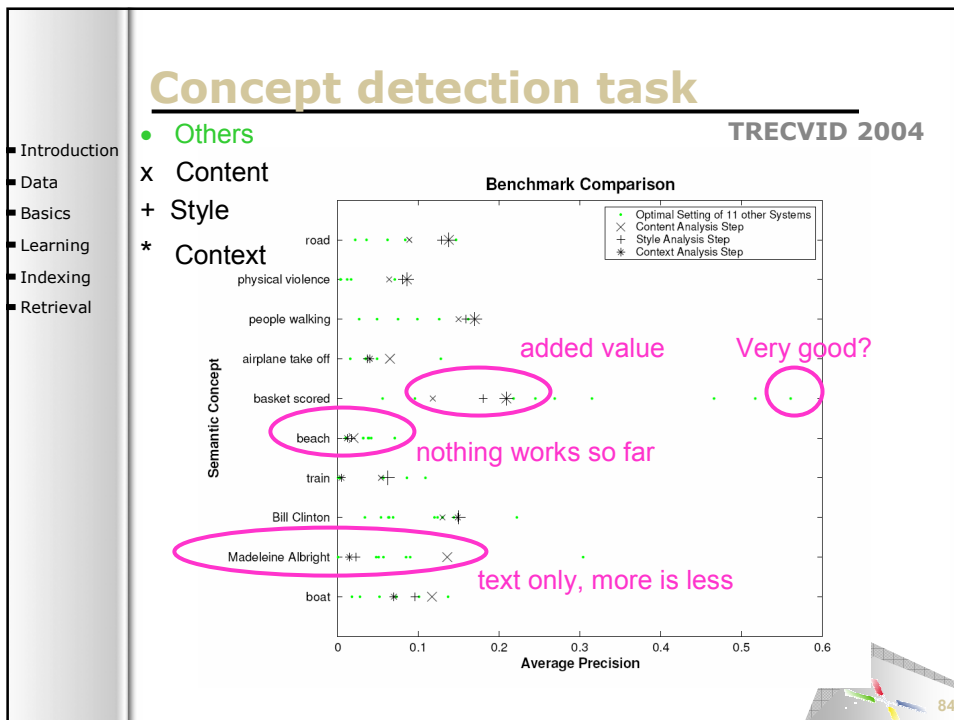
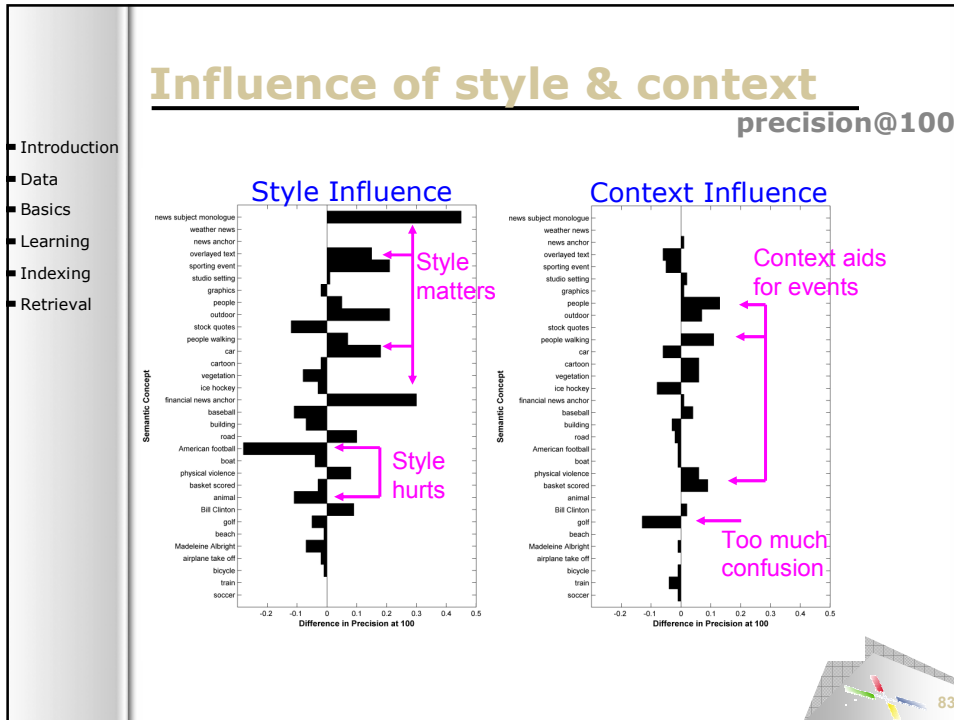
81

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	0.08	0.08	0.99	0.99
Overlaid text	0.84	0.99	0.99	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	0.76	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	0.72	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	0.47	0.47	0.54
Building	0.53	0.46	0.43	0.53
Road	0.45	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

82



- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Content analysis pathfinder

TRECVID 2005

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

85

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Content analysis pathfinder

TRECVID 2005

- Vary unimodal and multimodal combinations

Data flow conventions

- Multimedia data
- ➡ Feature vector
- ➡ Annotated concept lexicon
- - - - - Ranked concept detection result

Details in Wednesday's talk @ 08.30

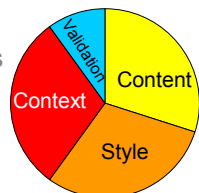

86

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Preliminaries

TRECVID 2005

- **Data preparation**
 - ✓ We randomly split training set a priori into 4 sets
 - ✓ Three sets for training (30%), 1 set for validation (10%)
- **Concept annotation**
 - ✓ TRECVID common annotation effort as basis
 - ✓ Extended manually to 101 concepts
 - ✓ Incomplete, but reliable
- **Machine learning architecture**
 - ✓ Support Vector Machine
 - ✓ Learn optimal parameters
 - ✓ Using 3 x 3-fold cross validation
 - ✓ Or grid-search on a 'grid'

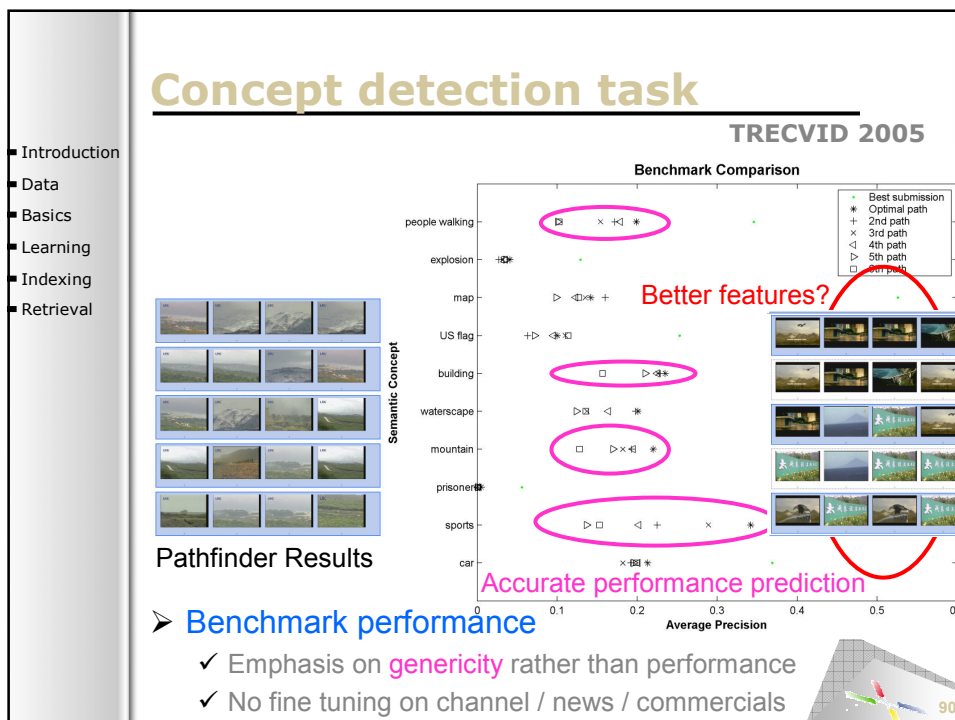
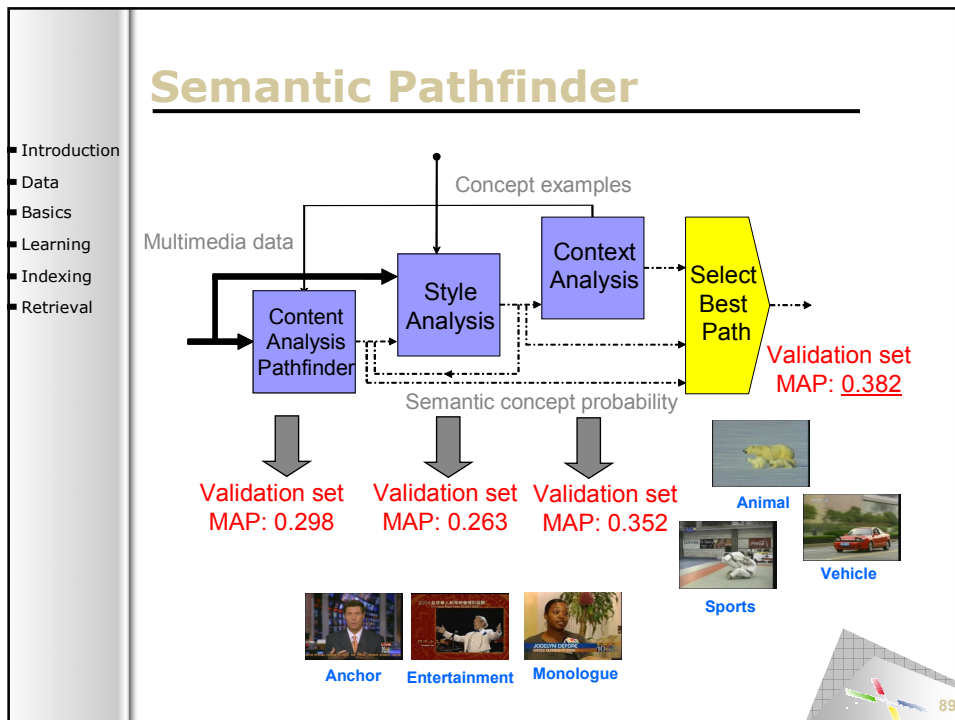
- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

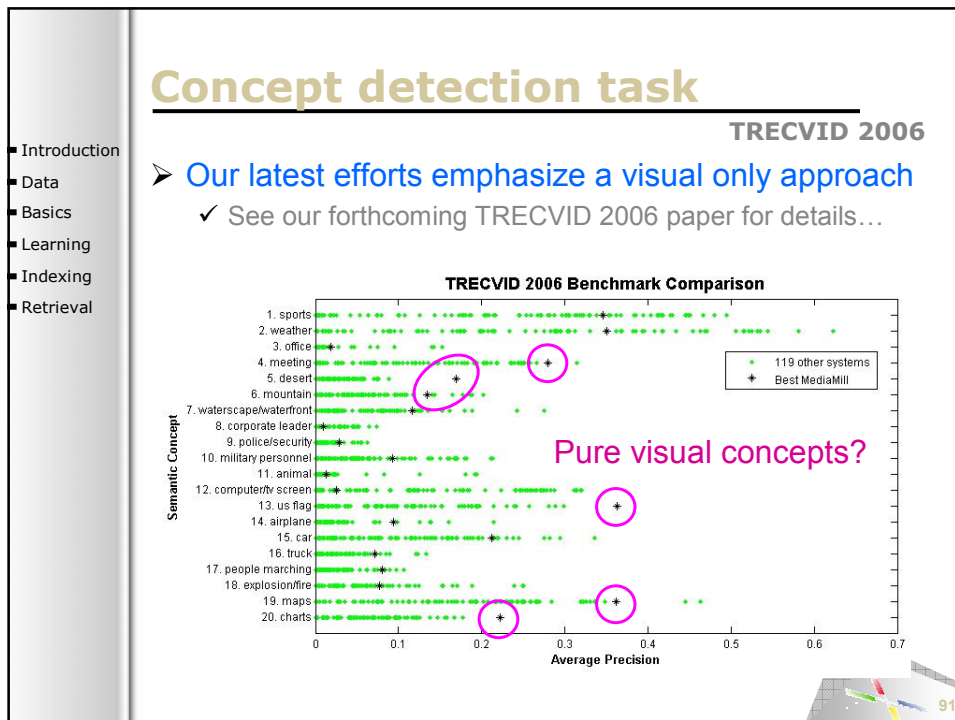
Annotated 101 concept lexicon

TRECVID 2005

Aircraft	I. Allawi	Anchor	Animal	Y. Arafat	Baseball	Basketball	Beach	Bicycle	Bird	T. Blair	Boat
Building	Bus	G. Bush Jr.	G. Bush sr.	Candle	Car	Cartoon	Chair	Charts	B. Clinton	Cloud	Corp. leader
Court	Crowd	Cycling	Desert	Dog	Drawing	Drawing & Cartoon	Duo-anchor	Entertainment	Explosion	Face	Female
Fire weapon	Fish	Flag	Flag USA	Food	Football	Golf	Government building	Government leader	Graphics	Grass	Horse
Horse racing	House	Indoor	H. Jintao	J. Kerry	E. Lshoud	Male	Map	Meeting	Military	Monologue	Motorbike
Mountain	H. Nasrallah	Natural disaster	News paper	Night fire	Office	Outdoor	Overlaid text	People	People marching	People walking	Police security
C. Powell	Prisoner	Racing	Religious leader	River	Road	Screen	A. Sharon	Sky	Smoke	Snow	Soccer
Split screen	Sports	Studio	Swimming pool	Table	Tank	Tennis	Tower	Tree	Truck	Urban	Vegetation
Vehicle	Violence	Waterfall	Waterscape	Weather							







Case study
Fabchannel.com

FABCHANNEL
PARADISO AND MELKWEG CONCERTS ONLINE

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

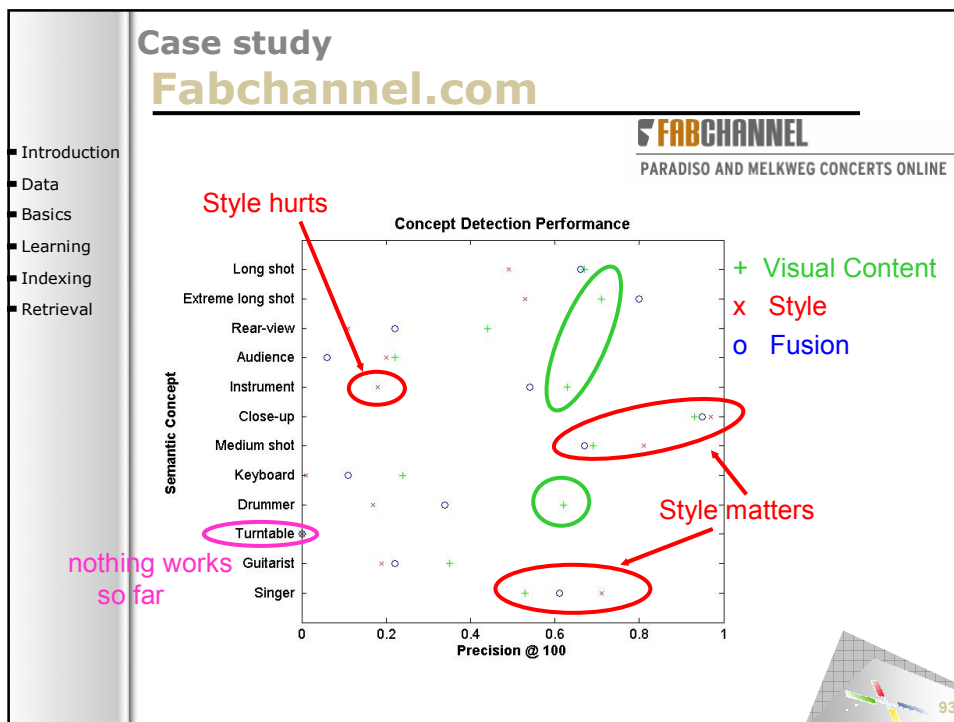
➤ Fabchannel request
✓ What can you do with 45 hours of live concerts?

➤ Answer:
✓ Let's try the semantic pathfinder

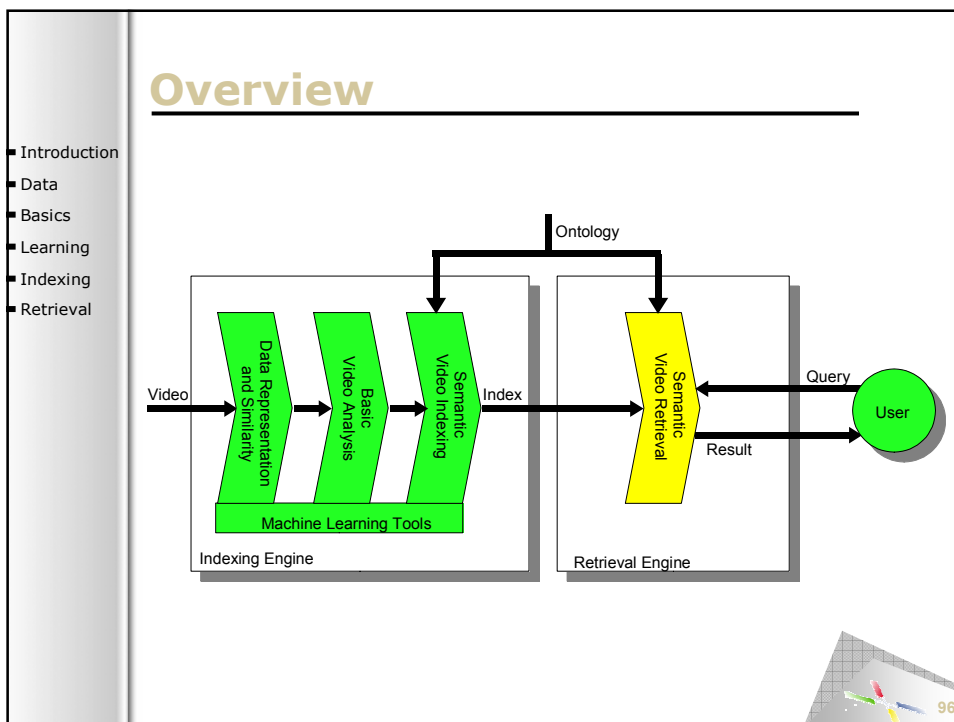
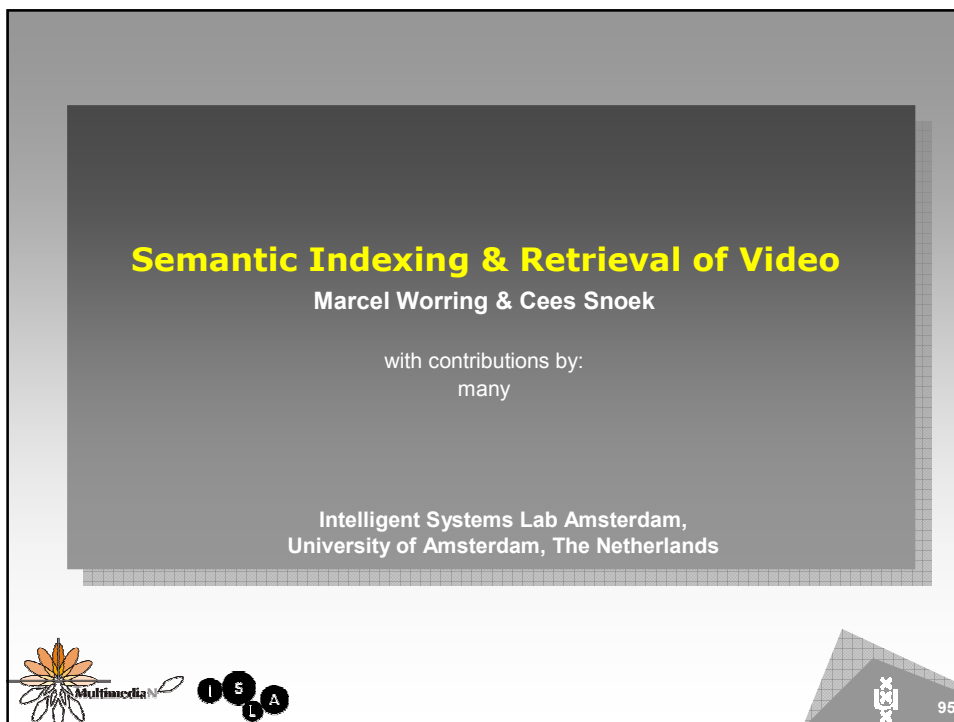
Extreme long shot Guitarist Instrument Singer Turntable Audience

Close-up Rear-view Long shot Drummer Keyboard Medium shot

92



- Conclusions**
- **Semantic pathfinder = generic video indexing**
 - ✓ Confirms the authoring metaphor
 - ✓ Currently detects up to ~~1~~ 450 concepts
 - **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



Introduction

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

1951 1992 1995 2000 yesterday tomorrow

← Semantic Gap →

Prior art

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Data flow conventions

- Multimedia raw data
- Feature vector
- Similarity distance

	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

Lessons learned

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

- **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

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

High-dimensional space for all words in speech transcript

Latent Semantic Indexing

Visual Similarity indexing

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Matrix containing all similarities between pairs of shots

101

Story level indexing

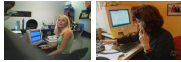
References:
IBM

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

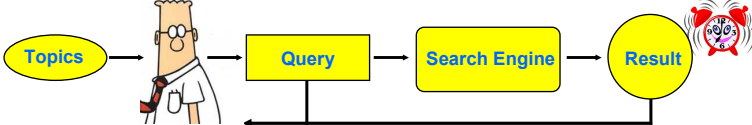
102

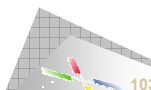
TRECVID interactive retrieval

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval



Find shots of an office setting






103

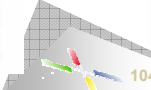
References:
FxPal

Browsers: MediaMagic

➤ Focus on the story level

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval





104

References:
Oulu University

Cluster-temporal browsing

➤ Using that result are typically similar and/or close in time

105

References:
IBM

IBM MARVeI

➤ A web based system


106

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

IBM MARVeI



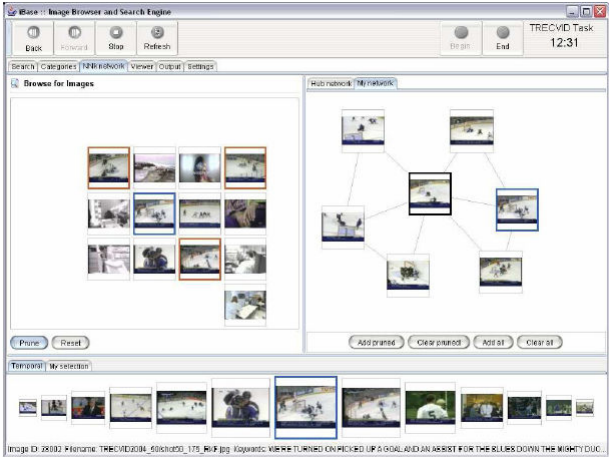
http://mp7.watson.ibm.com/marvel/

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

References:
Imperial College London

NN^k Browser

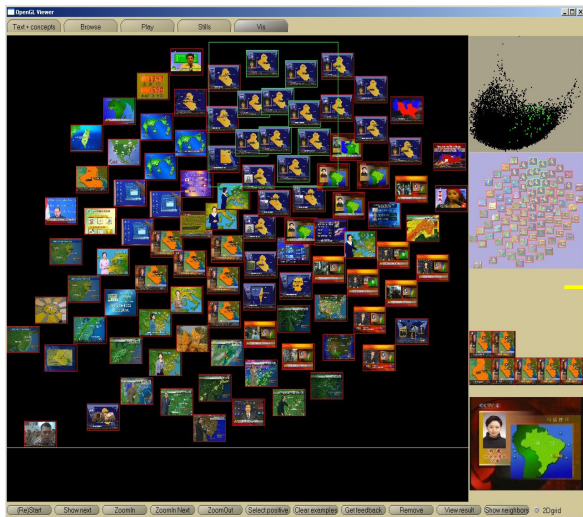
➤ Analyze the context of the current shot



- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

The GalaxyBrowser

➤ Pure similarity based browsing



Induced by similarity

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Extreme video retrieval

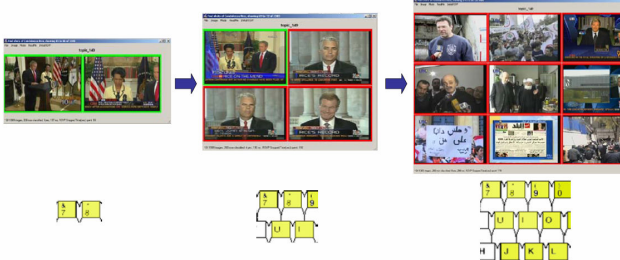
by Carnegie Mellon University

➤ **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

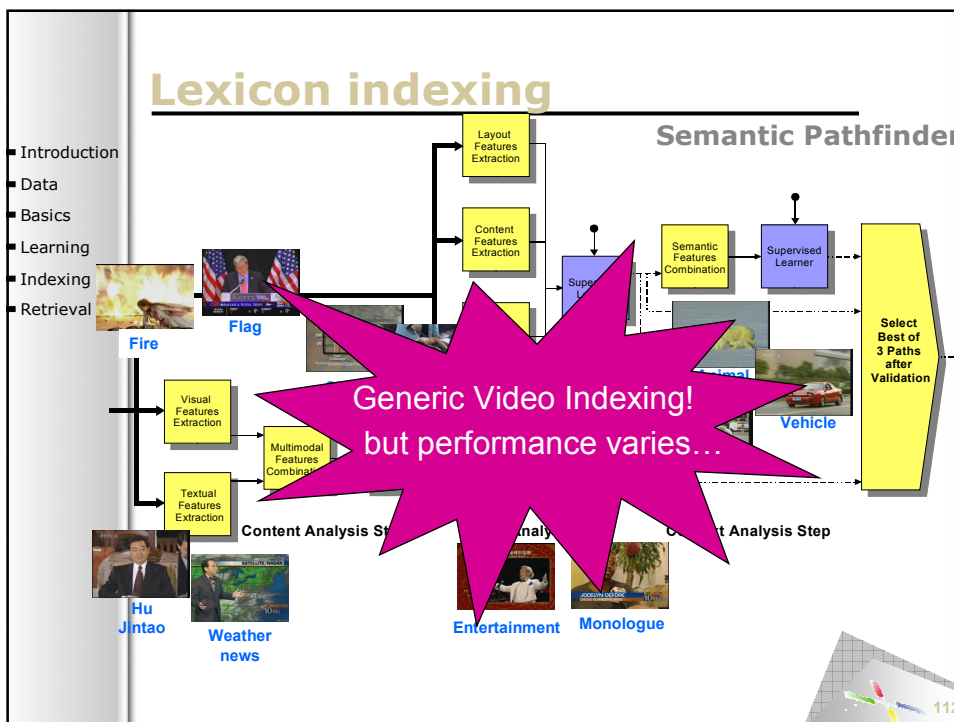


References:
Dublin City University

Físchlár

➤ Optimized for use by “real” users

The screenshot shows a web-based search interface. On the left is a 'QUERY PANEL' with a search bar and several filter categories (e.g., 'COLOR', 'SCENE', 'OBJECT'). The main area is a 'SEARCH RESULT' grid displaying multiple video thumbnails. On the right is a 'SAVED POINT' sidebar with a vertical list of thumbnails. The interface is clean and user-friendly.



- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Query selection

... yields a ranking of the data

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

Display of results

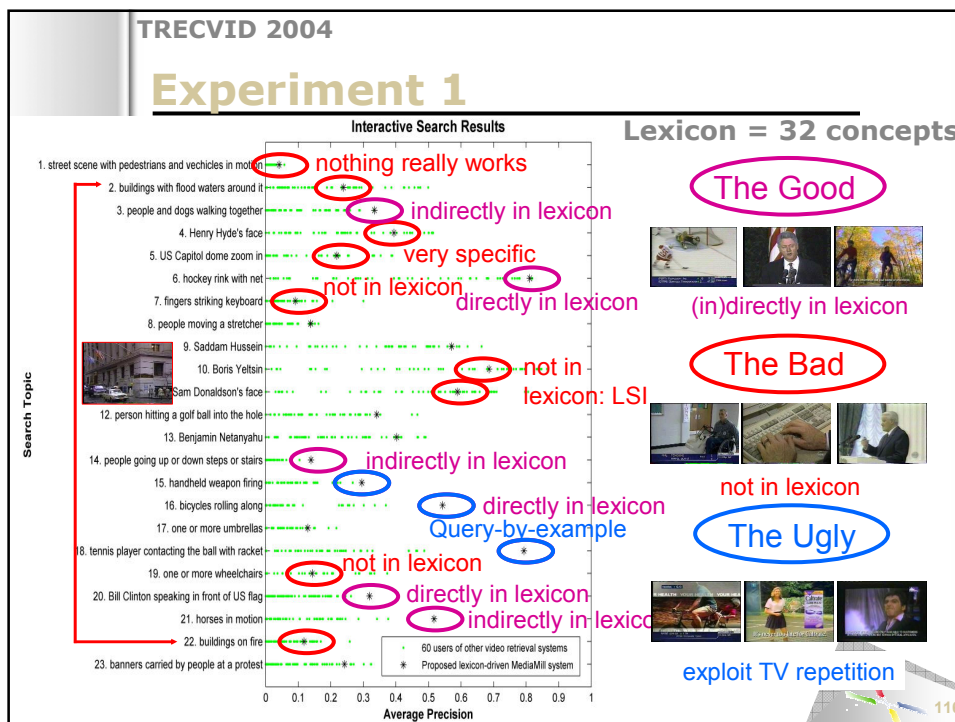
CrossBrowser

TRECVID 2004

Learned lexicon of 32 concepts

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

115



TRECVID 2005

Learned lexicon of 101 concepts

117

TRECVID 2005

Experiment 2

Lexicon = 101 concepts

Interactive Search Results

1. Condoleezza Rice (not in lexicon)

2. Iyad Allawi (poor concept detection)

3. Omar Karami

4. Hu Jintao

5. Tony Blair

6. Mahmoud Abbas

7. graphic map of Iraq, Baghdad marked (concept specification fails)

8. two visible tennis players on the court

9. people shaking hands

10. helicopter in flight

11. George W. Bush entering or leaving a vehicle

12. something on fire with flames and smoke

13. people with banners or signs

14. people entering or leaving a building

15. a meeting with a large table and people (concept combination fails)

16. a ship or boat

17. basketball players on the court

18. one or more palm trees

19. an airplane taking off

20. a road with one or more cars

21. one or more military vehicles

22. a tall building

23. a goal being made in a soccer match

24. office setting

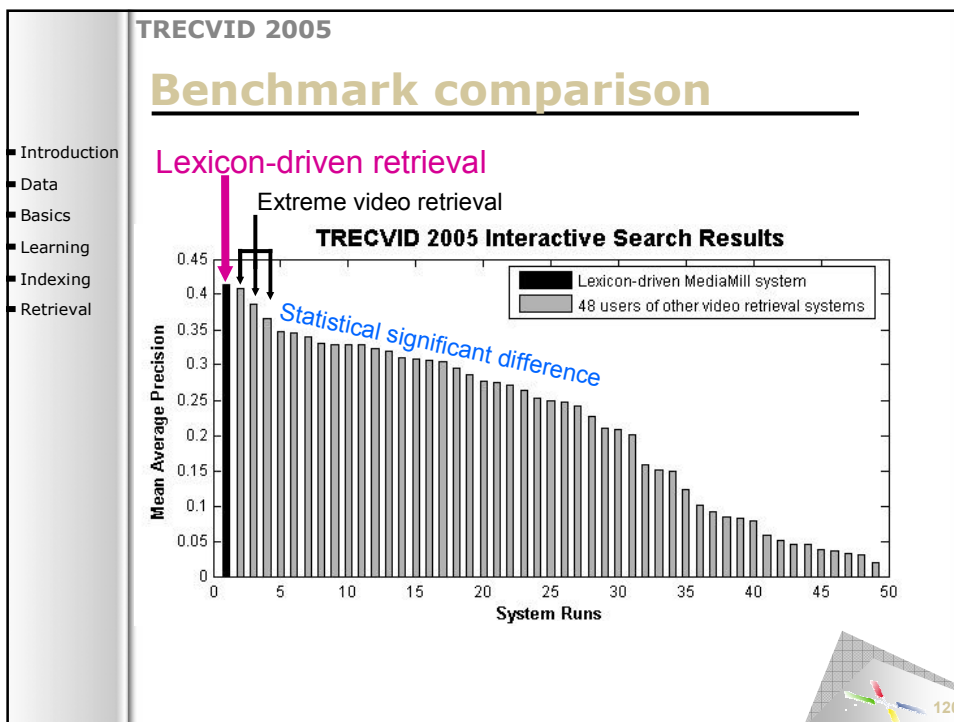
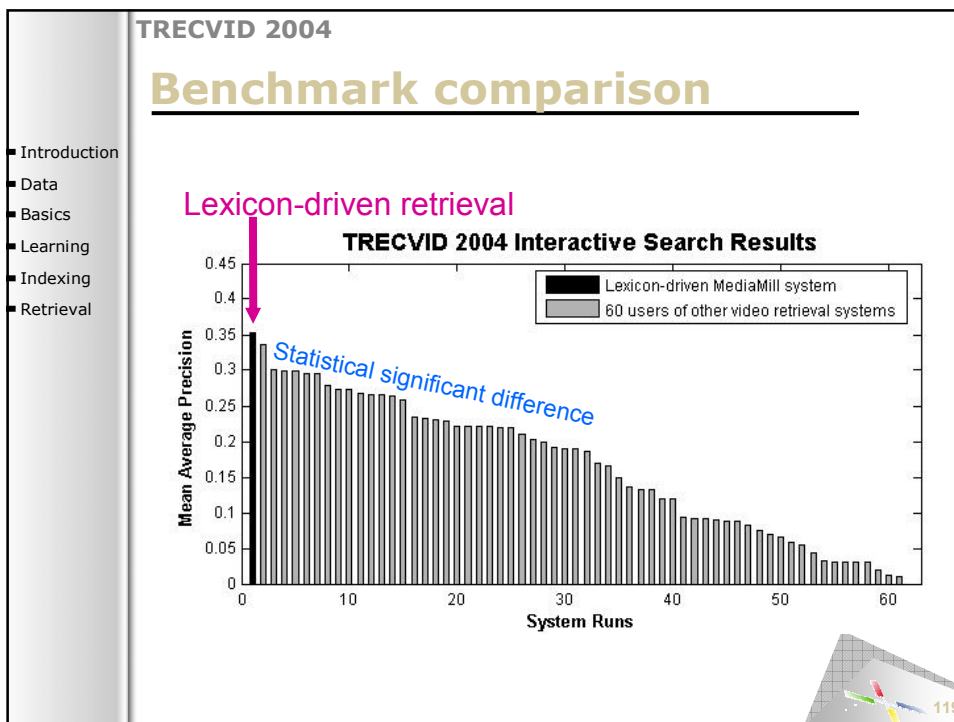
Legend: 48 users of other video retrieval systems (dots), Proposed lexicon-driven MediaMill system (asterisks)

The Good
Almost all topics solvable by using concept lexicon only!

The Bad

The Beautiful
Exploit common sense!

118



Use wordnet relationships

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

```
graph TD; Vehicle -- "Is a kind of" --> Bus; Vehicle -- "Is a kind of" --> Bicycle; Vehicle -- "Is a kind of" --> Car; Vehicle -- "Is a kind of" --> Tank;
```

123

Ontology querying

References:
Bouke Huurnink (UvA) & Laura Hollink (VU)

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

1. Identify objects in WordNet

124

References:
Bouke Huurnink (UvA) & Laura Hollink (VU)

Ontology querying

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

2. Identify related concept detectors

125

References:
Bouke Huurnink (UvA) & Laura Hollink (VU)

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

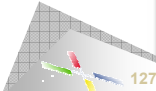
126

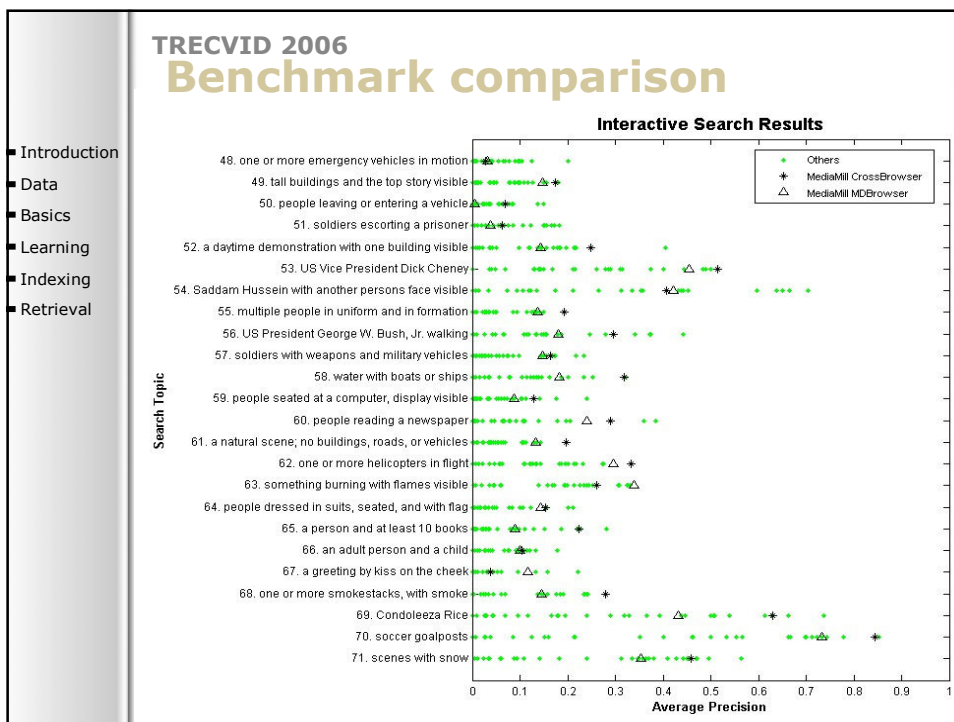
Multi concept browsing

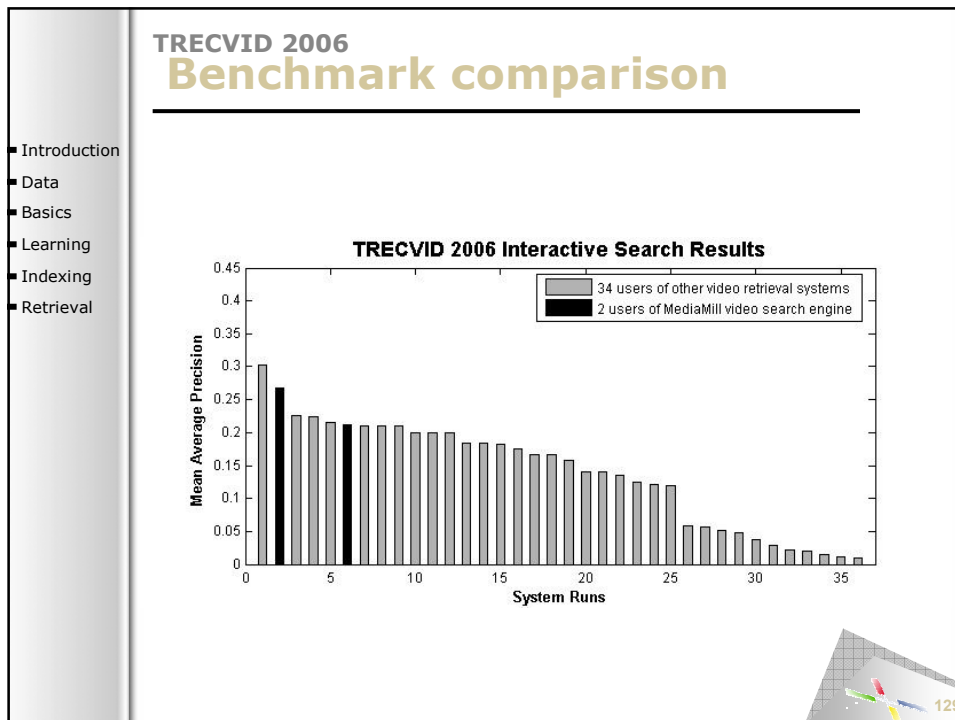
But when there are many relevant concept I want to be able to browse through different dimensions



- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval







Further information

- Introduction
- Data
- Basics
- Learning
- Indexing
- Retrieval

➤ Including the sheets of the tutorial

www.mediamill.nl

And come see our demo on Wednesday

130