Semantic Video Indexing

T-61.6030 Multimedia Retrieval

Stevan Keraudy
stevan.keraudy@tkk.fi

Helsinki University of Technology
March 14, 2008

What is it?

- Query by keyword or tag is common
- Semantic Video Indexing aims at:
  - Analyzing the content of a video
  - Recognizing some concepts
  - Indexing the video depending on concepts
- It uses machine learning techniques to learn the concepts

Background

- Ch.2: Metadata
- Ch.3: Pattern recognition
- Ch.4,5,7: Unimodal media analysis

Outline

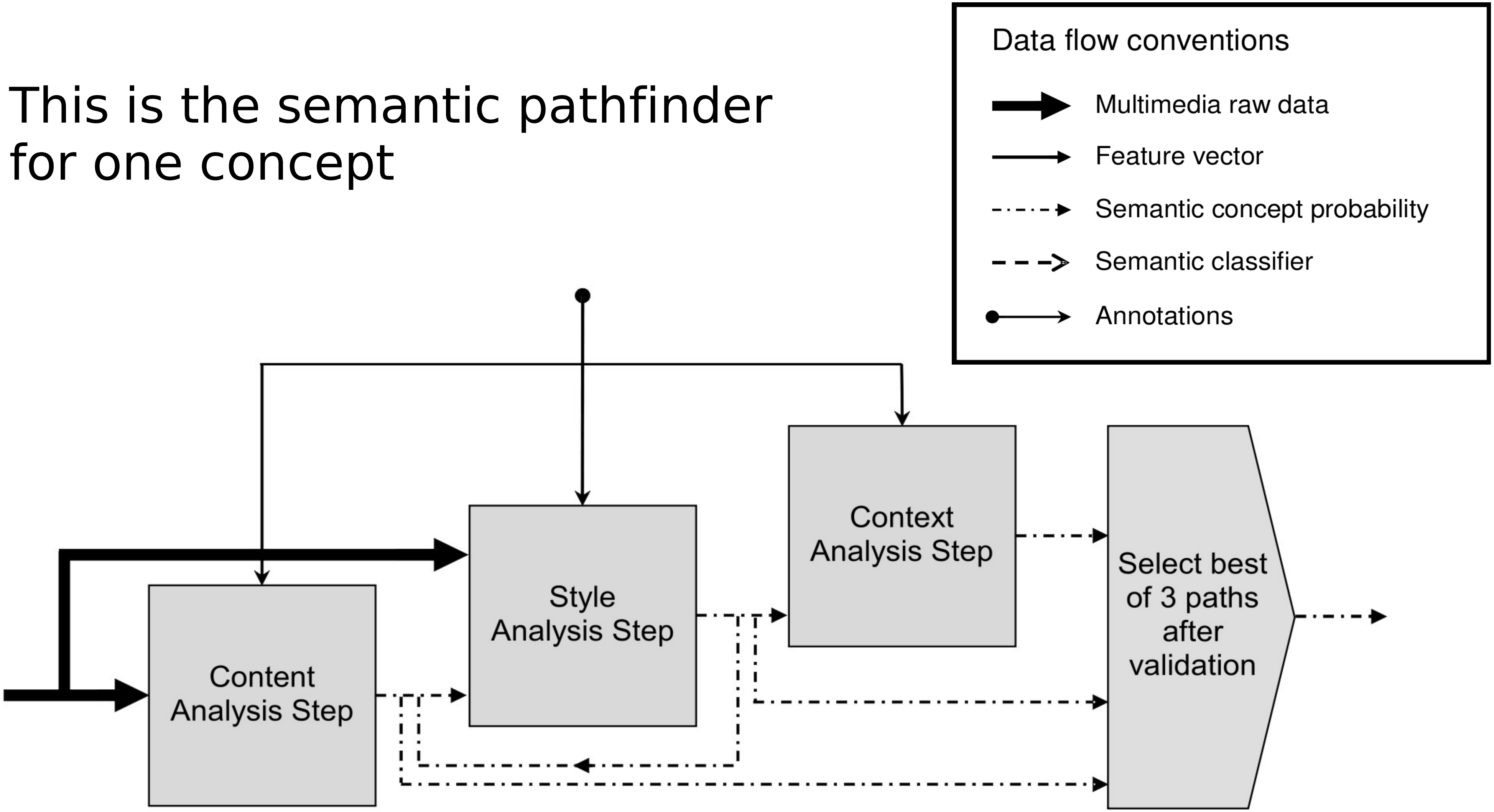
- Semantic pathfinder
    - Description
    - Content analysis step
    - Style analysis step
    - Context analysis step
    - Semantic pathfinder output
- Experiments on real-world data
    - Description
    - Results analysis

Semantic Pathfinder

- 3 consecutive analysis steps:
  - Content analysis step
  - Style analysis step
  - Context analysis step
- One step's output can be used for next one's input
- Depending on the concept, we might want to ignore some steps

Semantic Pathfinder

This is the semantic pathfinder for one concept



Data Set

- Focused on news video
- 184 hours of ABC and CNN news
- MPEG-1 format
- Training set: 120h (Jan. 98 – Jun. 98)
- Test set: 64h (Oct. 98 – Dec. 98)
- Analysis tries to recognize 32 concepts in this data set

Concept Lexicon

Airplane take off

American football

Animal

Baseball

Basket scored

Beach

Bicycle

Bill Clinton

Boat

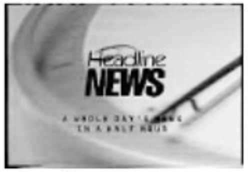
Building

Car

Cartoon

Financial news anchor

Golf



Graphics

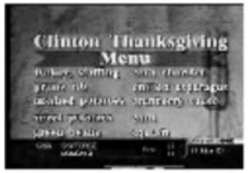
Ice hockey

Madeleine Albright

News anchor

News subject Monologue

Outdoor



Overlayed text

People

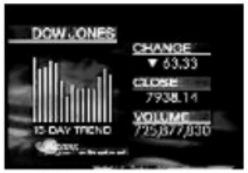
People walking

Physical violence

Road

Soccer

Sporting event

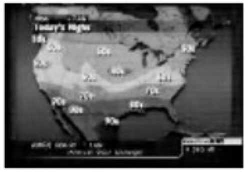


Stock quotes

Studio setting

Train

Vegetation


Weather news

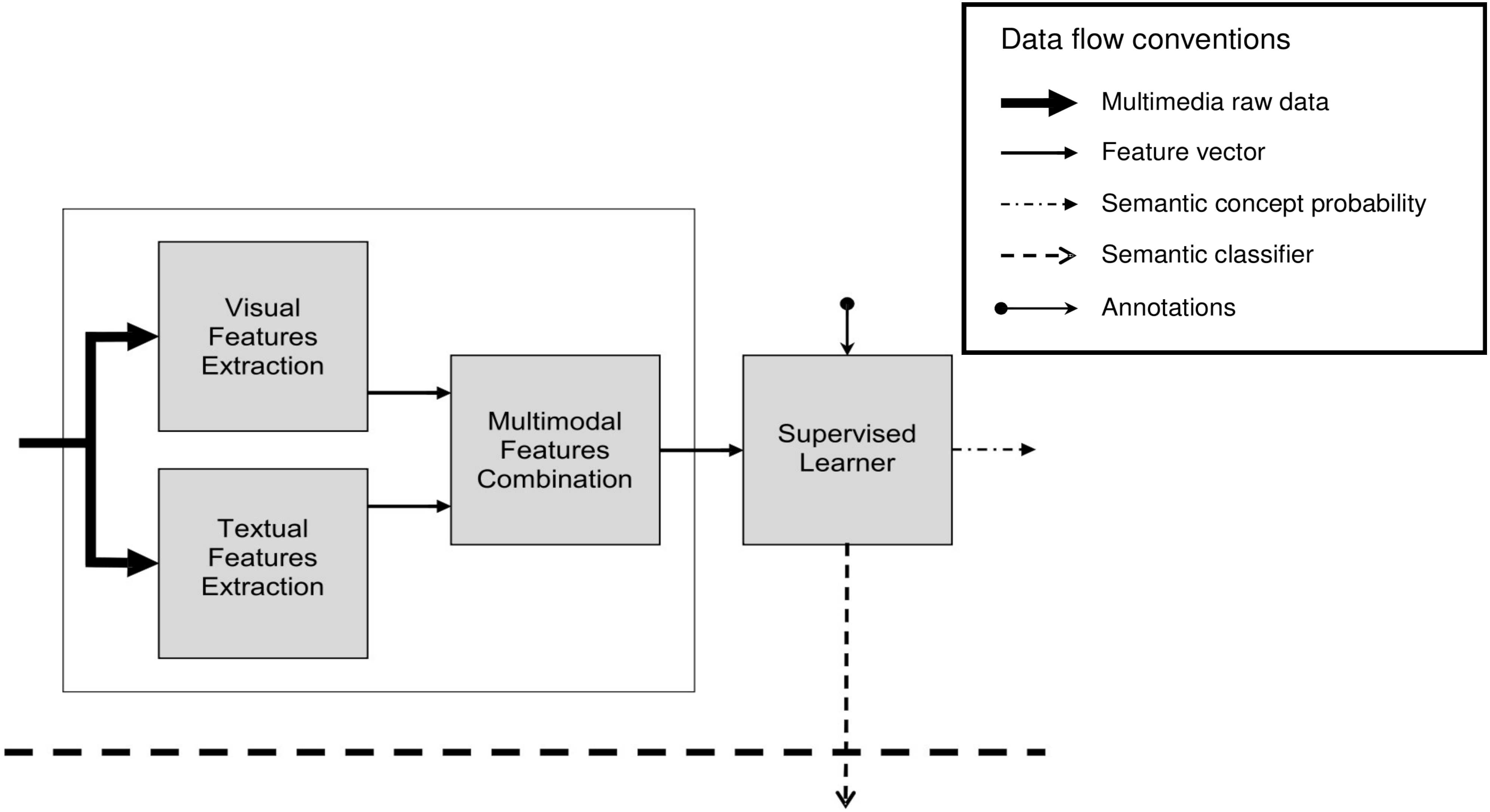
Analysis Step Architecture

- Semantic video indexing is a pattern recognition problem
  - Segment a video
  - Select relevant shots
  - Given pattern *x*, detect semantic concept *w* from shot *i*
  - Each step extracts *x_i* and learns *p(w|x_i)* for each concept *w*
- Support Vector Machine is used

Content Analysis Step

- 3 sub-steps:
  - Visual analysis: extract visual features
  - Text analysis: extract speech transcript
  - Multimodal analysis: combine both features

Content Analysis Step


Data flow conventions
Multimedia raw data
Feature vector
Semantic concept probability
Semantic classifier
Annotations
Visual Features Extraction
Textual Features Extraction
Multimodal Features Combination
Supervised Learner

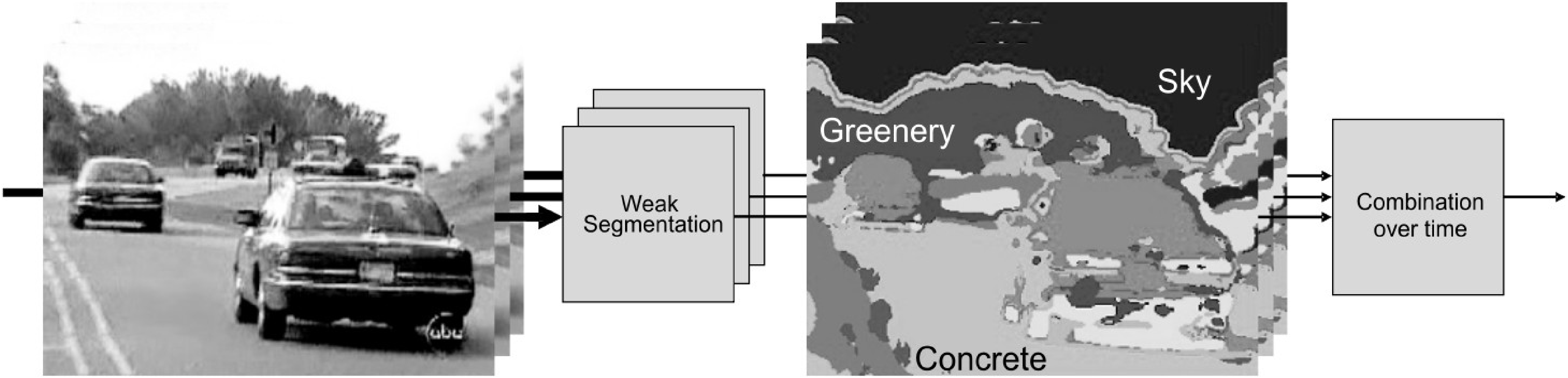
Visual Analysis

- Regional visual concepts:
  - {colored clothing, concrete, fire, grassland, greenery, red carpet, sand, sky,...}
- Segmentation of each image frame using color invariance
- Invariant features extraction is computed for each pixel of each frame
- We use SVM to classify each pixel

Visual Analysis

- A combination over time is made
- We select one frame out of a sequence that represents the best the features
- This choice is made by an SVM
- The output is an image vector
    - For each regional visual concept, it indicates the percentage of pixels of this class

Visual Analysis

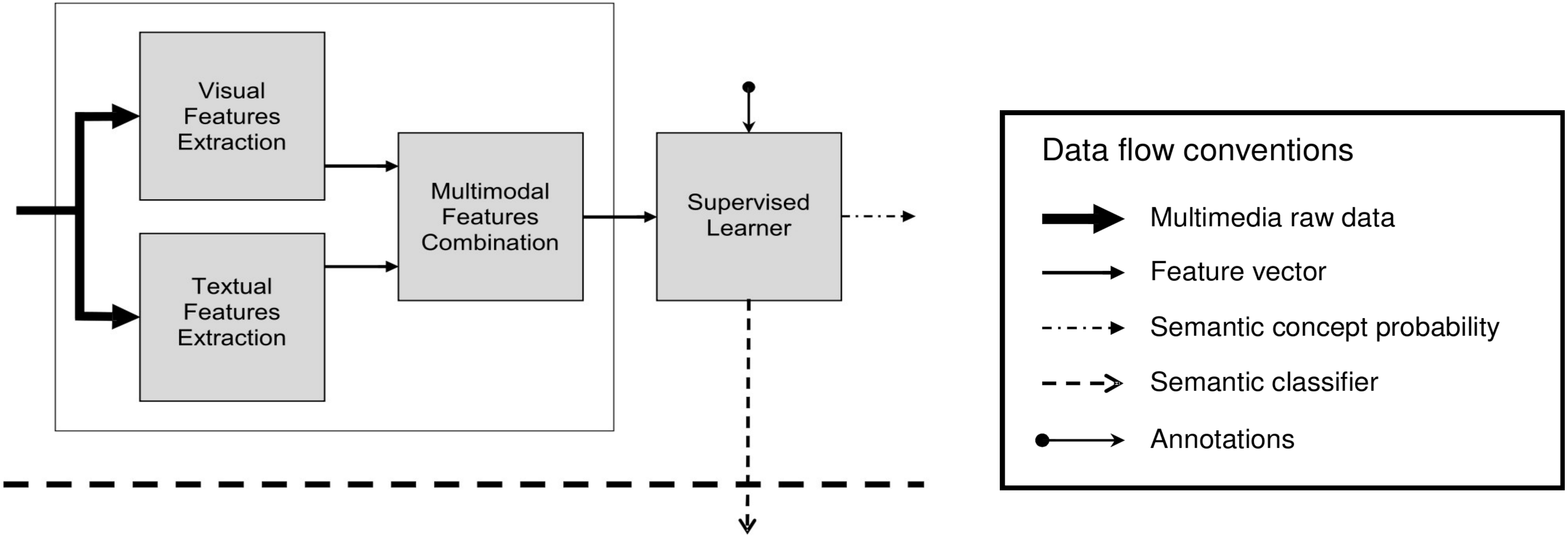


Textual Analysis

- Speech is transcribed into text
- Stop-words removal
- A new lexicon is created for every concept using training data
- We compare the text associated with shots with the lexicons
- Special treatment for *Persons* concept

Multimodal Analysis

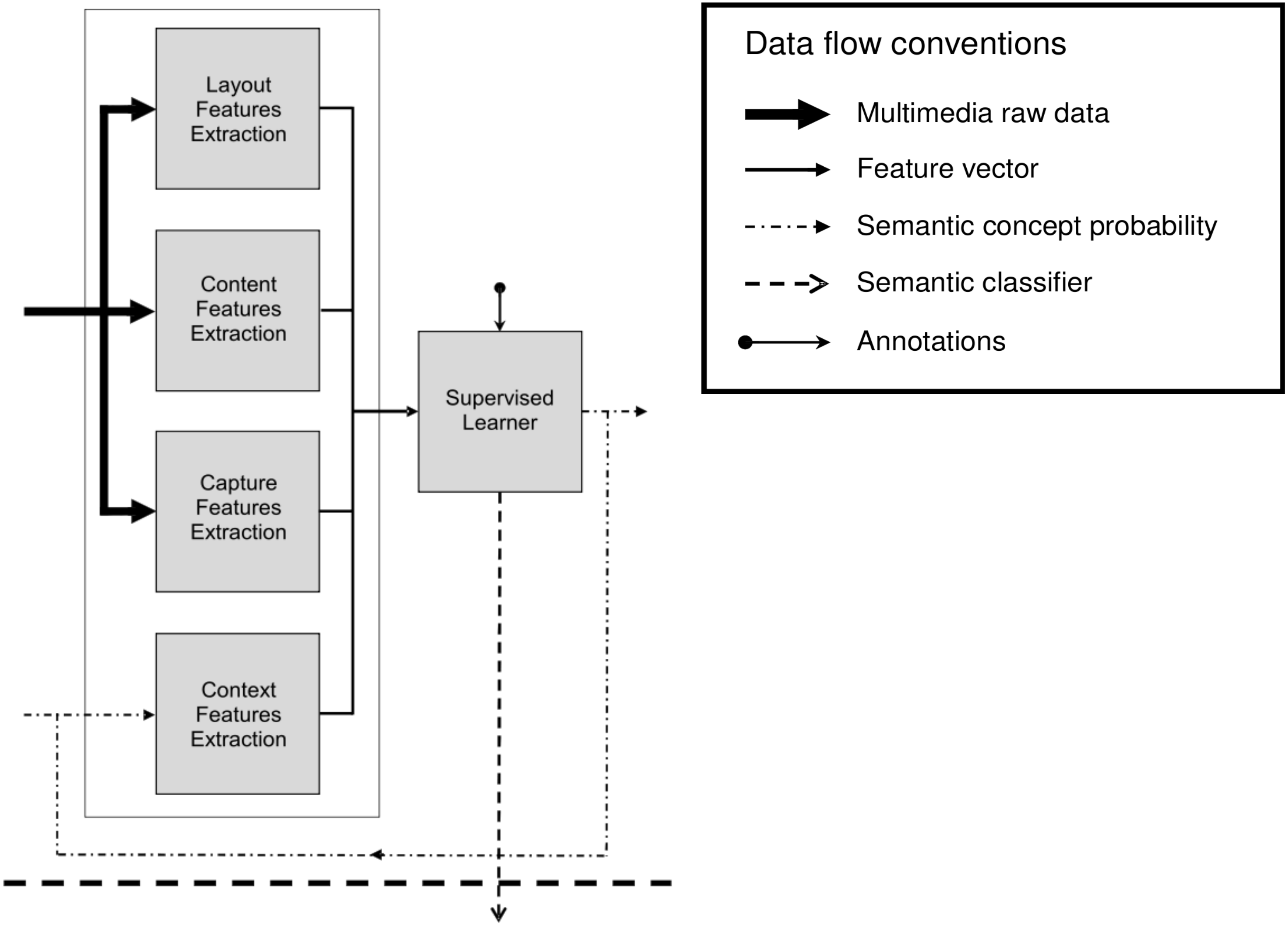
- We concatenate visual analysis and textual analysis outputs
- We feed the supervised learning module


Visual Features Extraction
Textual Features Extraction
Multimodal Features Combination
Supervised Learner
Data flow conventions
Multimedia raw data
Feature vector
Semantic concept probability
Semantic classifier
Annotations

Style Analysis Step

- Video is viewed from a production perspective
- 4 production roles are detected by different algorithms
- Feature extraction is independent of the data set
- Then we use an iterative classification

Style Analysis Step



Layout Features Extraction
Content Features Extraction
Capture Features Extraction
Context Features Extraction
Supervised Learner

Data flow conventions

- Multimedia raw data
- Feature vector
- Semantic concept probability
- Semantic classifier
- Annotations

Layout

- 4 features are used:
  - Shot length
  - Overlayed text
  - Silence
  - Voice-over

Content

- 8 features are used:
  - Faces (frontal face detector)
  - Face location
  - Cars
  - Object motion
  - Frequent speaker (3 most frequent speakers)
  - Overlayed text length
  - Video text named entity
  - Voice named entity (from transcript)

Capture

- 3 features are used:
  - Camera distance (from size of faces)
  - Camera work (pan, tilt, zoom,...)
  - Camera motion

Context

- Enhance or reduce correlation between semantic concepts
- Reduces number of false positives
- Increases number of true positives
- E.g. co-occurence of space shuttle and bicycle is improbable
- It takes as input the output of content analysis step

Iterative Classification

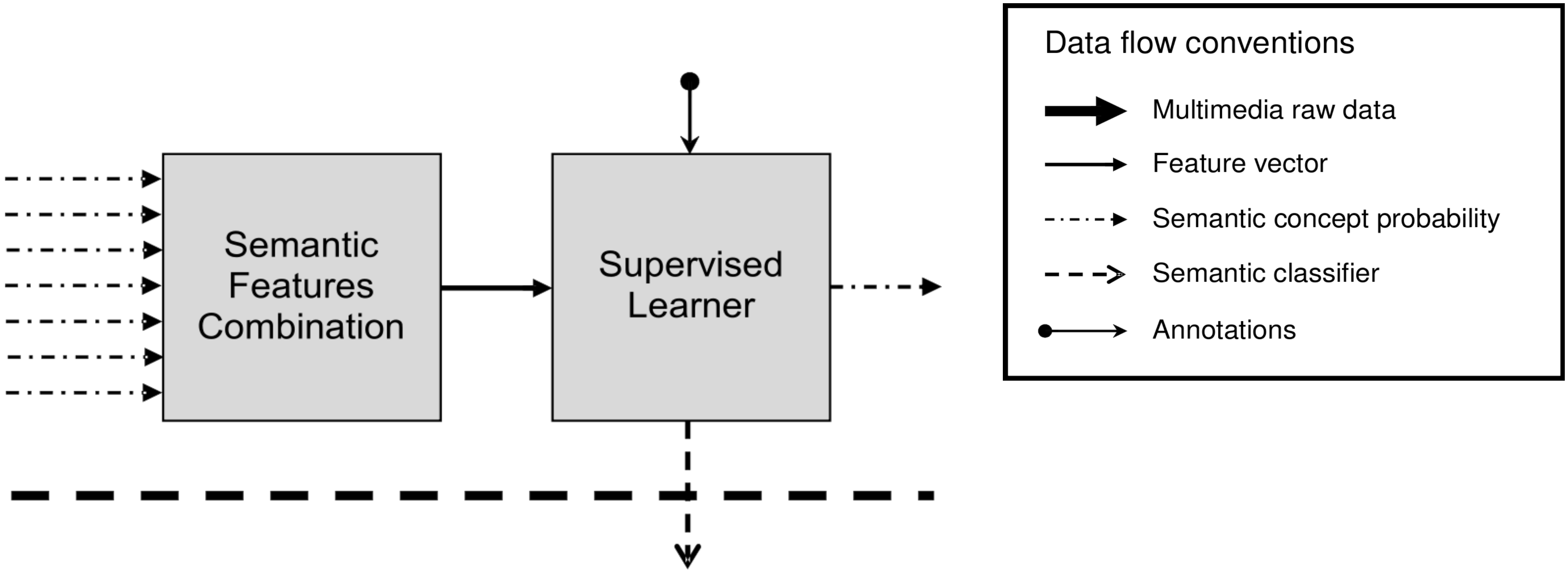
- For each concept *w* in lexicon:
  - Take as input the output of content analysis step and results of style analysis step
  - Classify
  - Update content analysis step output
- The output of the whole iteration serves as an input to next analysis step

Style Analysis Step Output

- For each shot *i*
  - For each concepts *w* in lexicon
    - Return $p(w|i)$
- Output is made of all probabilities $p(w|i)$

Context Analysis Step

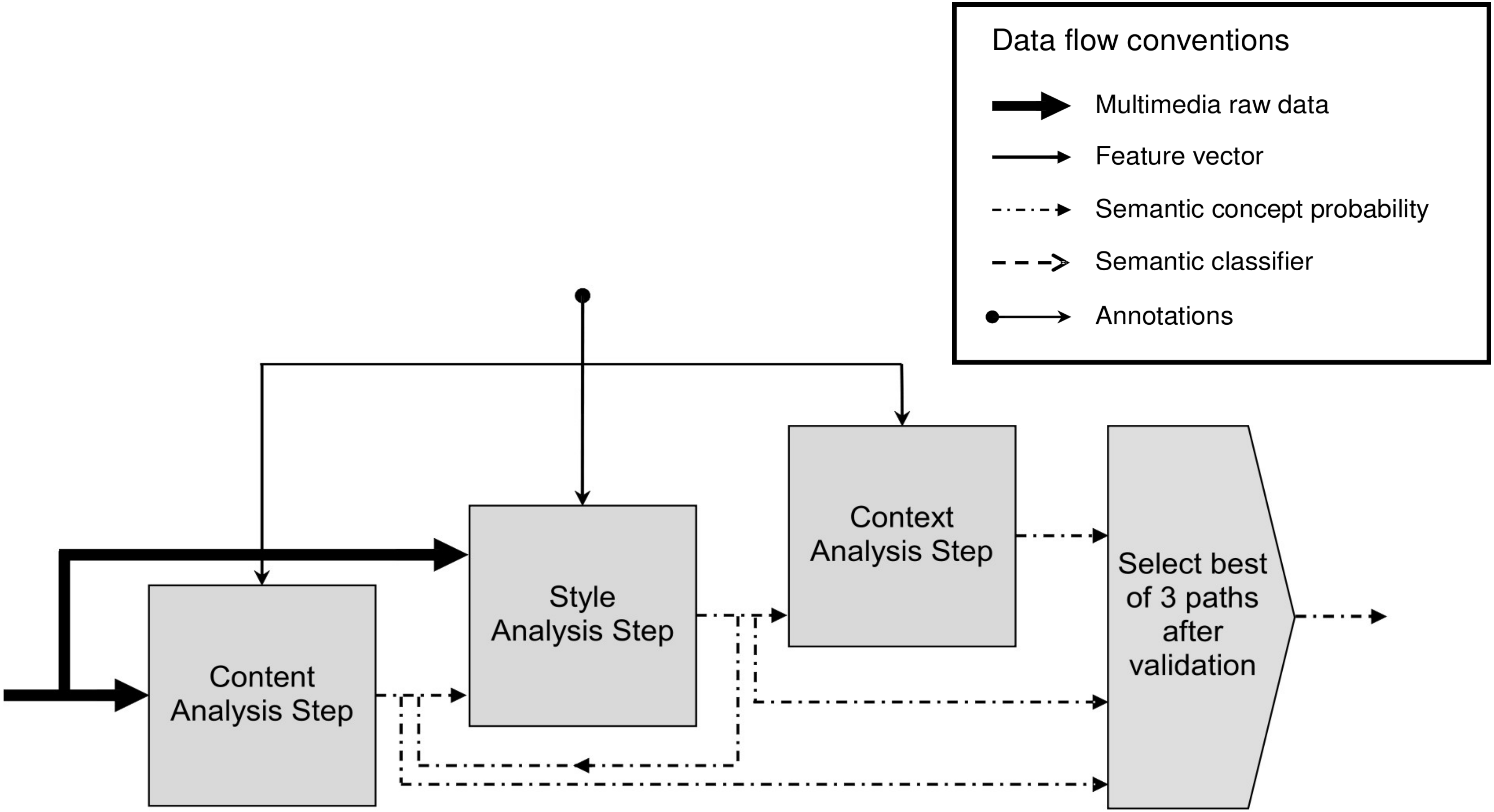
- Takes as input all concepts probabilities
- Learns relations between concepts



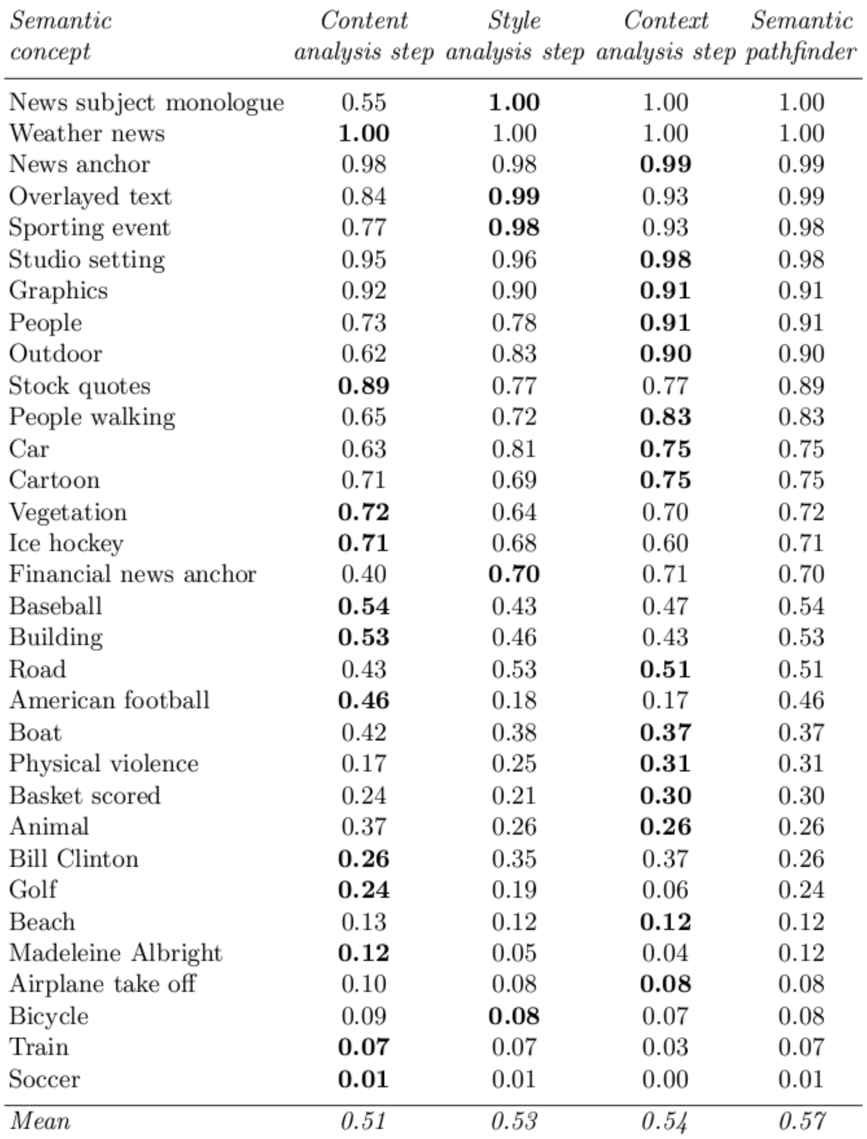
Semantic Pathfinder Output

- Output of context analysis step gives pathfinder global output
- For each concept w we get:
  - p(w|content)
  - p(w|content, style)
  - p(w|content, style, context)
- We select one of these outputs for each concept.

Semantic Pathfinder Output



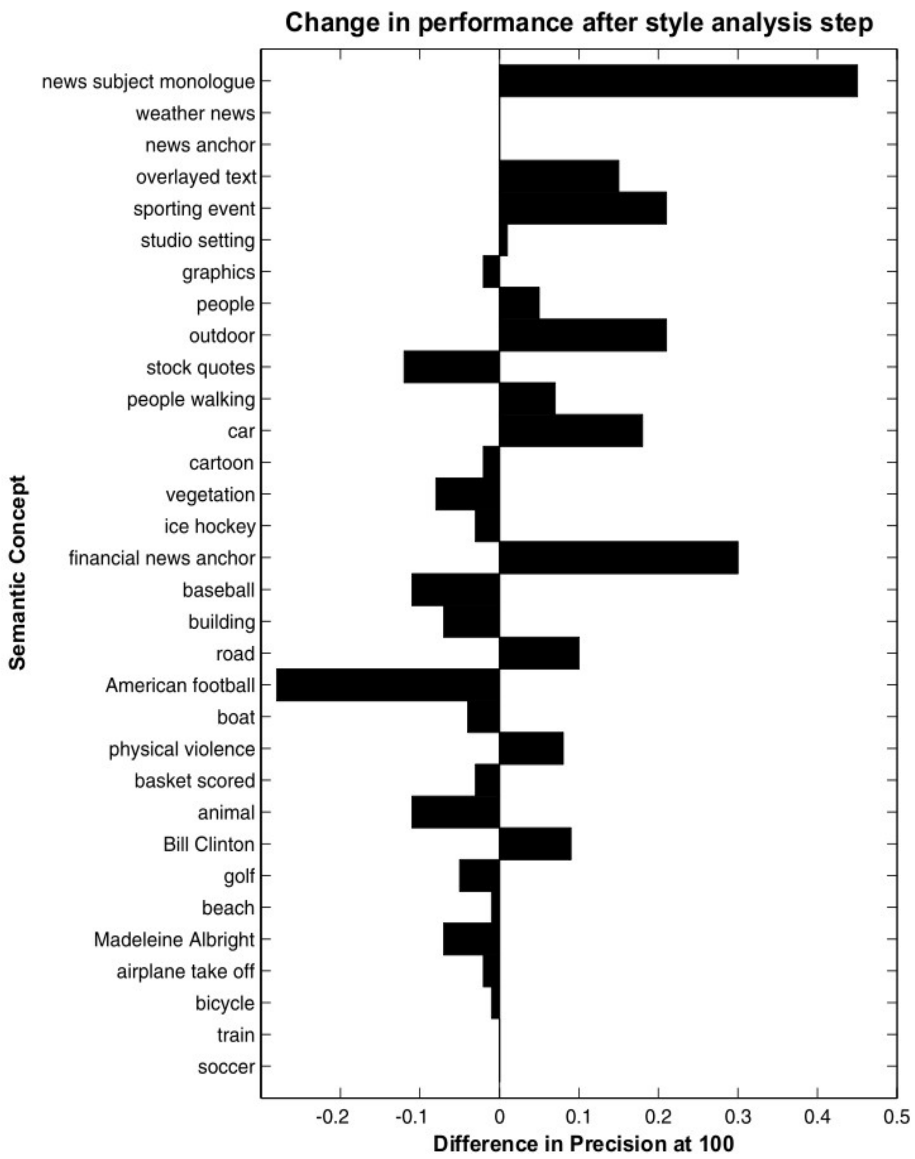
Experiments Results



Semantic concept	*Content analysis step*	*Style analysis step*	*Context analysis step*	*Semantic 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	0.43	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*

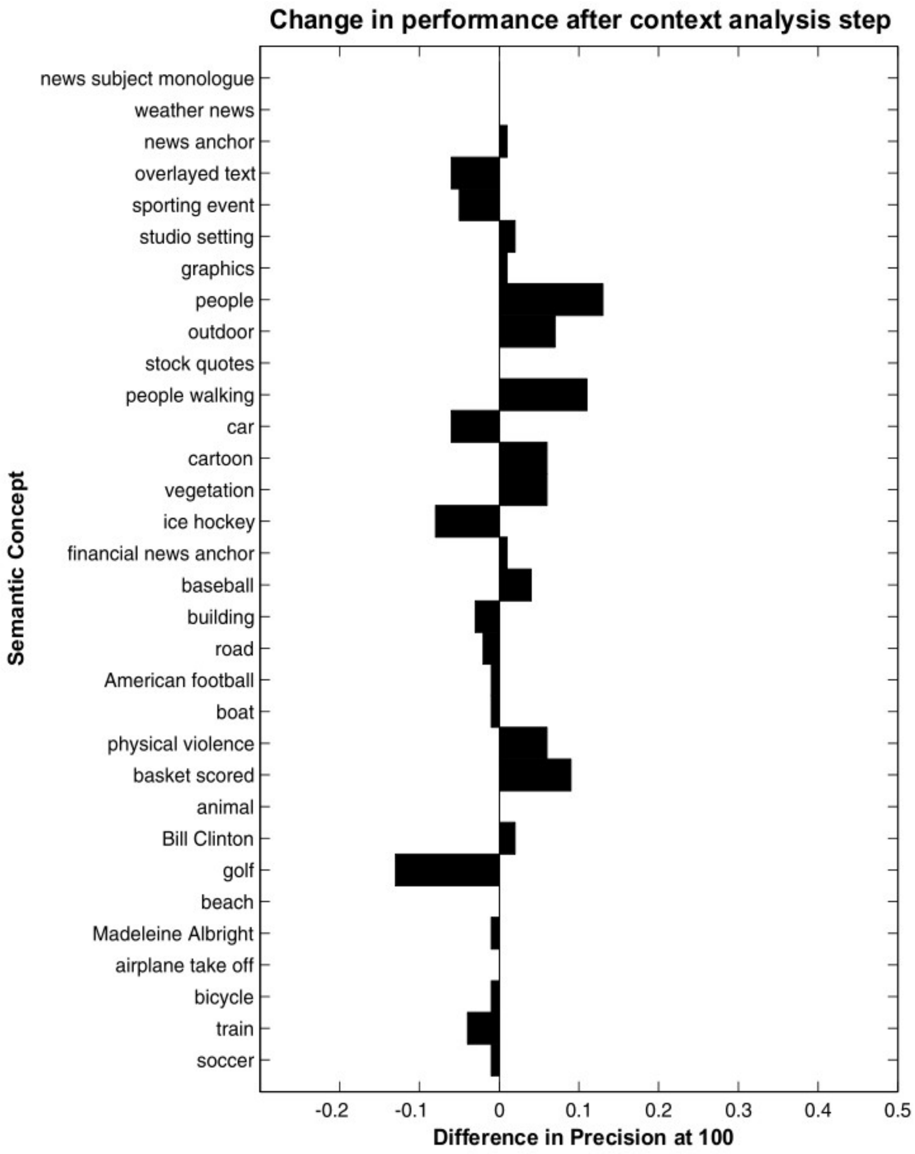
- 32 semantic concepts
- Precision is the percentage of correct shots
- Best results varies over concepts
  - Content: 12
  - Style: 5
  - Context: 15
- Global precision increases

Style Analysis Influence



- Increase for 12 concepts
- Especially semantically rich concepts

Context Analysis Influence



- Increase for 13 concepts
- *People* profits from sport-related concepts
- *Golf* suffers from *Outdoor* and *Vegetation*

Applications

- Semantic Video Search Engines
- Have a look at MediaMill
  - `http://www.science.uva.nl/research/mediamill/`
  - Query-by-concept using 32 concepts
  - Query-by-keyword
  - Query-by-example

References

- Blanken et al., *Multimedia Retrieval*, Springer, 2007 (Chapter 8)
- All images are also from this book