University of Marburg
Department of Mathematics and Computer Science
Distributed Systems Group

University of Siegen
SFB/FK 615
„Media Upheavals“

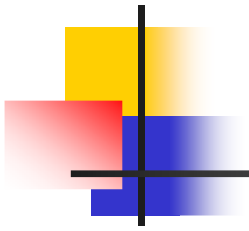
On the Generalization Capabilities of our High-Level Feature Detection System at TRECVID 2007

Markus Mühling[1,2], Ralph Ewerth[1,2], Thilo Stadelmann[1,2],

Christian Zöfel[2], Bing Shi[2], and Bernd Freisleben[1,2]

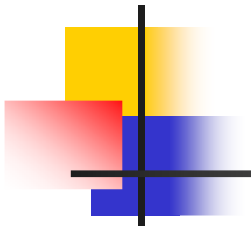
[1]SFB/FK 615, Research Project „Media Upheavals“, University of Siegen, Germany

[2]Dept. of Math. and Computer Science, University of Marburg, Germany



Outline

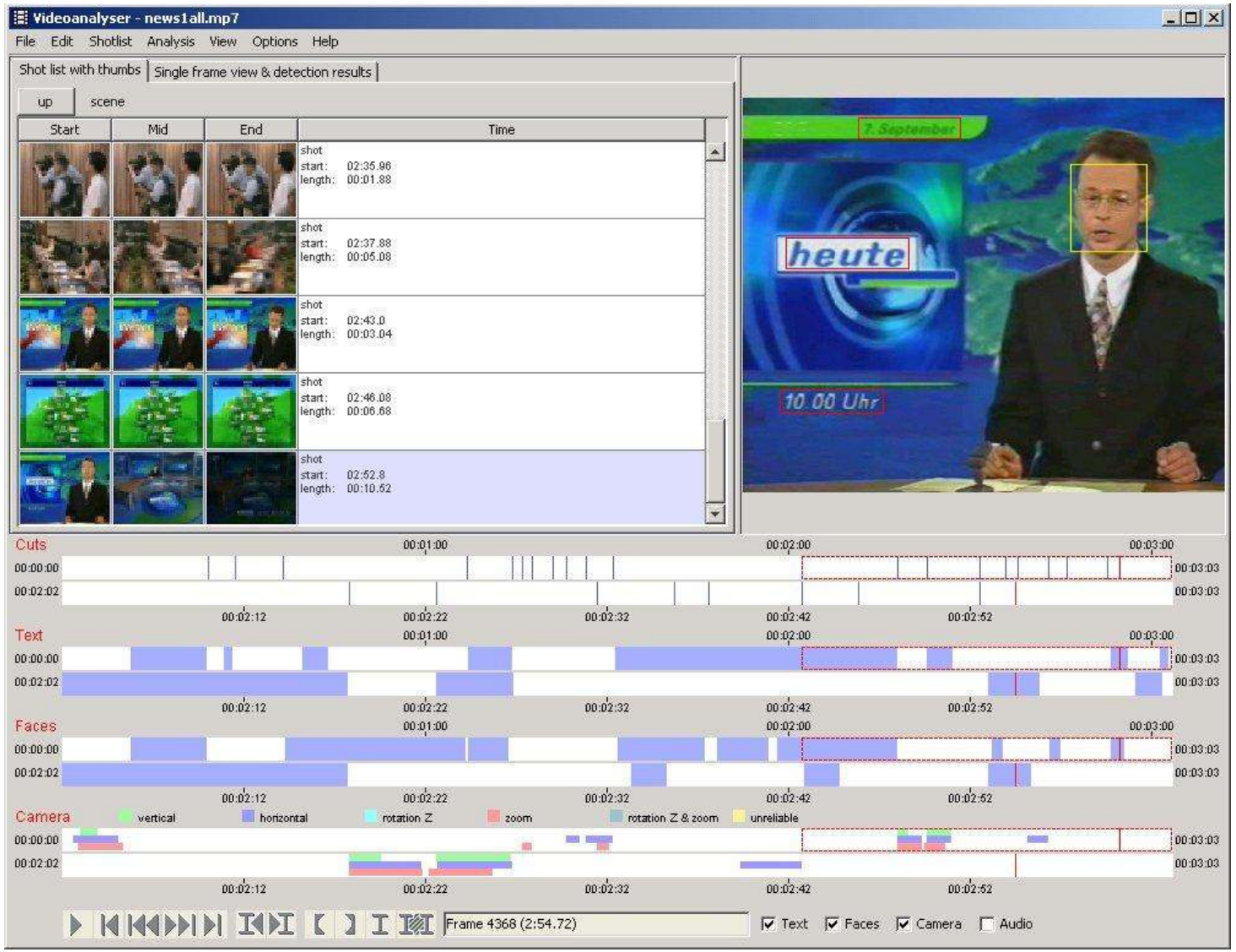
- Videana
- System overview
- Results
- Conclusions

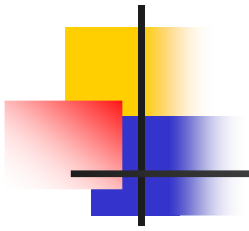


Videana

- Result of the research project
  "Methods and Tools for Computer-Assisted Media Analysis"
- Trans-disciplinary research
  - Support film studies in media sciences
- Goals
  - Computer-based methods for scientific analysis of images and videos
  - Development of a Grid-based infrastructure for
    - easy data access
    - parallel execution of compute-intensive algorithms
- Funded by Deutsche Forschungsgemeinschaft
  (DFG, SFB/FK 615), since 2002

Videana GUI

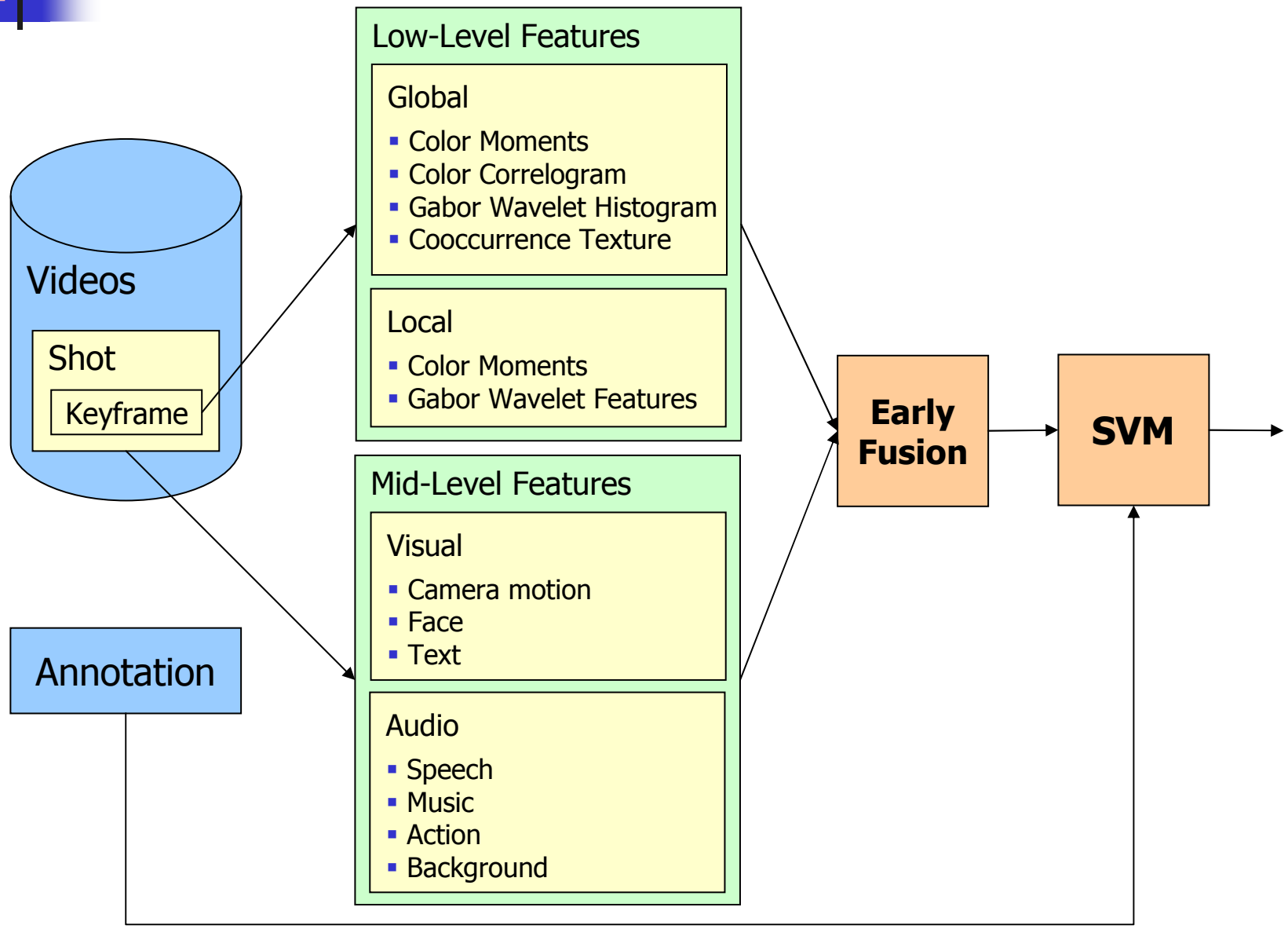

Videoanalyser - news1all.mp7
File Edit Shotlist Analysis View Options Help
Shot list with thumbs
Single frame view & detection results
up
scene
Start
Mid
End
Time
shot
start: 02:35.96
length: 00:01.88
shot
start: 02:37.88
length: 00:05.08
shot
start: 02:43.0
length: 00:03.04
shot
start: 02:46.08
length: 00:06.68
shot
start: 02:52.8
length: 00:10.52
heute
10.00 Uhr
Cuts
Text
Faces
Camera
vertical
horizontal
rotation Z
zoom
rotation Z & zoom
unreliable
Frame 4368 (2:54.72)
Text
Faces
Camera
Audio



Outline

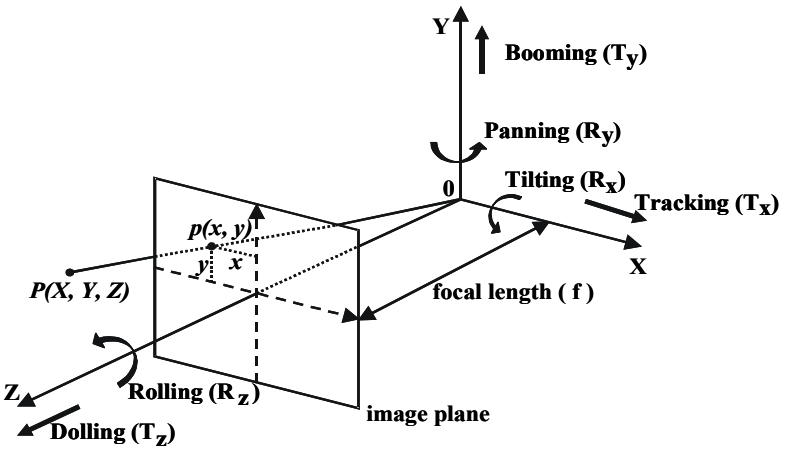
- Videana
- System overview
- Results
- Conclusions

Baseline System Overview



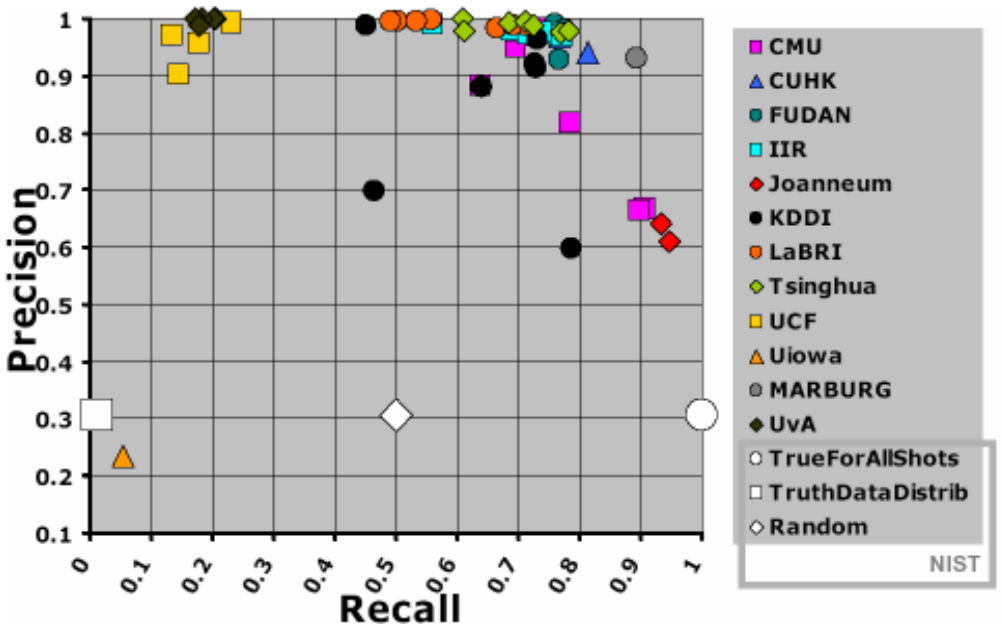
Camera Motion Estimation

1. Extraction of motion vectors from P-frames
2. Computing a reliable motion vector field
   - Apply outlier removal algorithm
3. Estimation of camera motion parameters
   - Apply Nelder-Mead minimization algorithm to estimate camera parameters of the 3D model

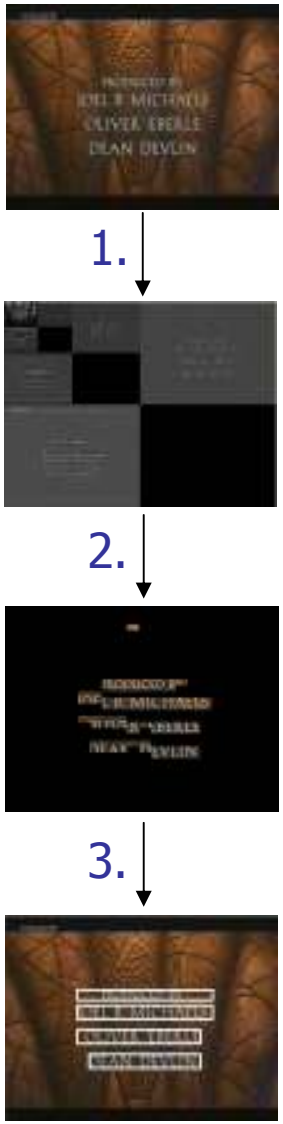


Camera Motion Estimation II

- Derived features per shot:
  - Percentage of pan, tilt respectively zoom
  - Statistical description of the distributions
- Very good results at TRECVID 2005
  - e.g. zoom:



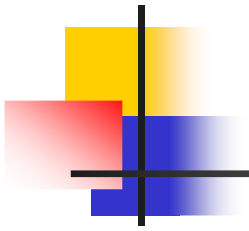
Text Detection and Localization



1. Wavelet transformation

2. Unsupervised text detection
   - Image regions represented by standard deviations of wavelet coefficients in different subbands
   - K-means clustering (text, background, complex background)

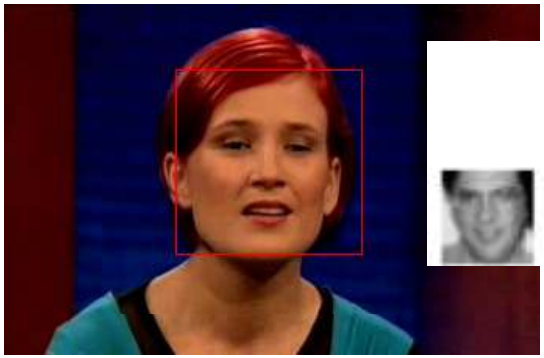
3. Text localization and refinement
   - Estimate connected components in the text cluster
   - Geometrical Analysis and refinement



Text Detection and Localization II

- Robust text detection approach
- Automatically detects horizontally aligned text with different sizes, fonts, colors and languages
- Derived features per shot
  - # text elements
  - Average text position
  - Text frame coverage
  - Average number of text elements per frame

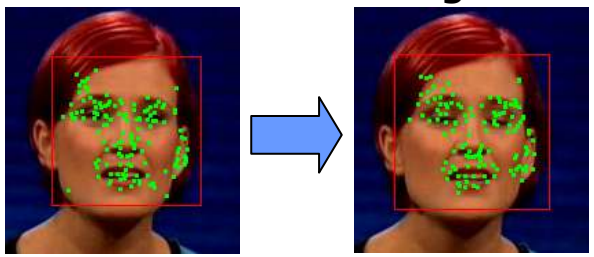
Face Processing



Frontal & Profile face detection
[Viola&Jones, 2004]

Skin color verification

Feature tracking



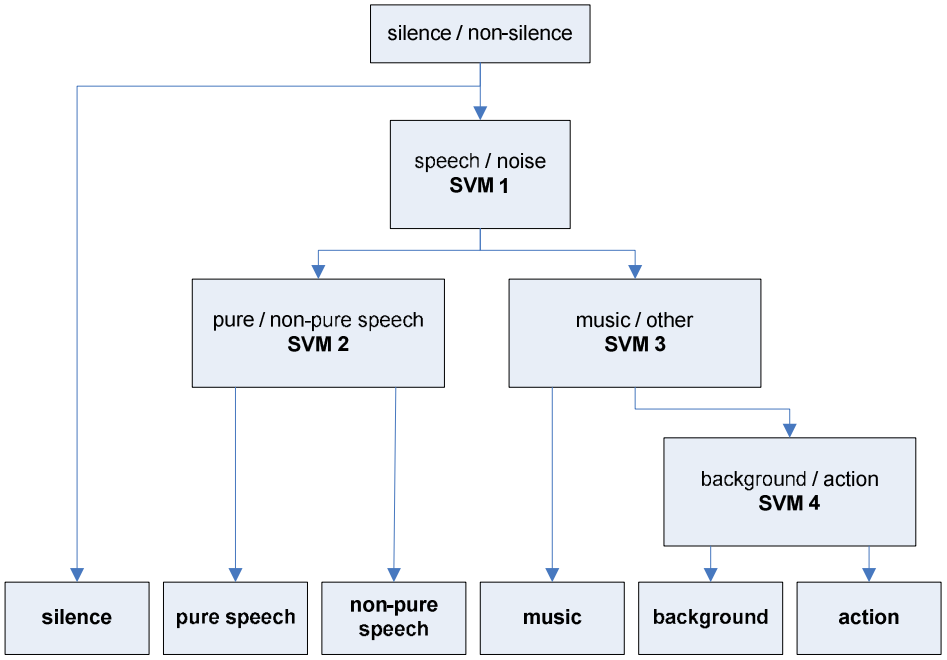
Face sequences

Derived mid-level features per shot:

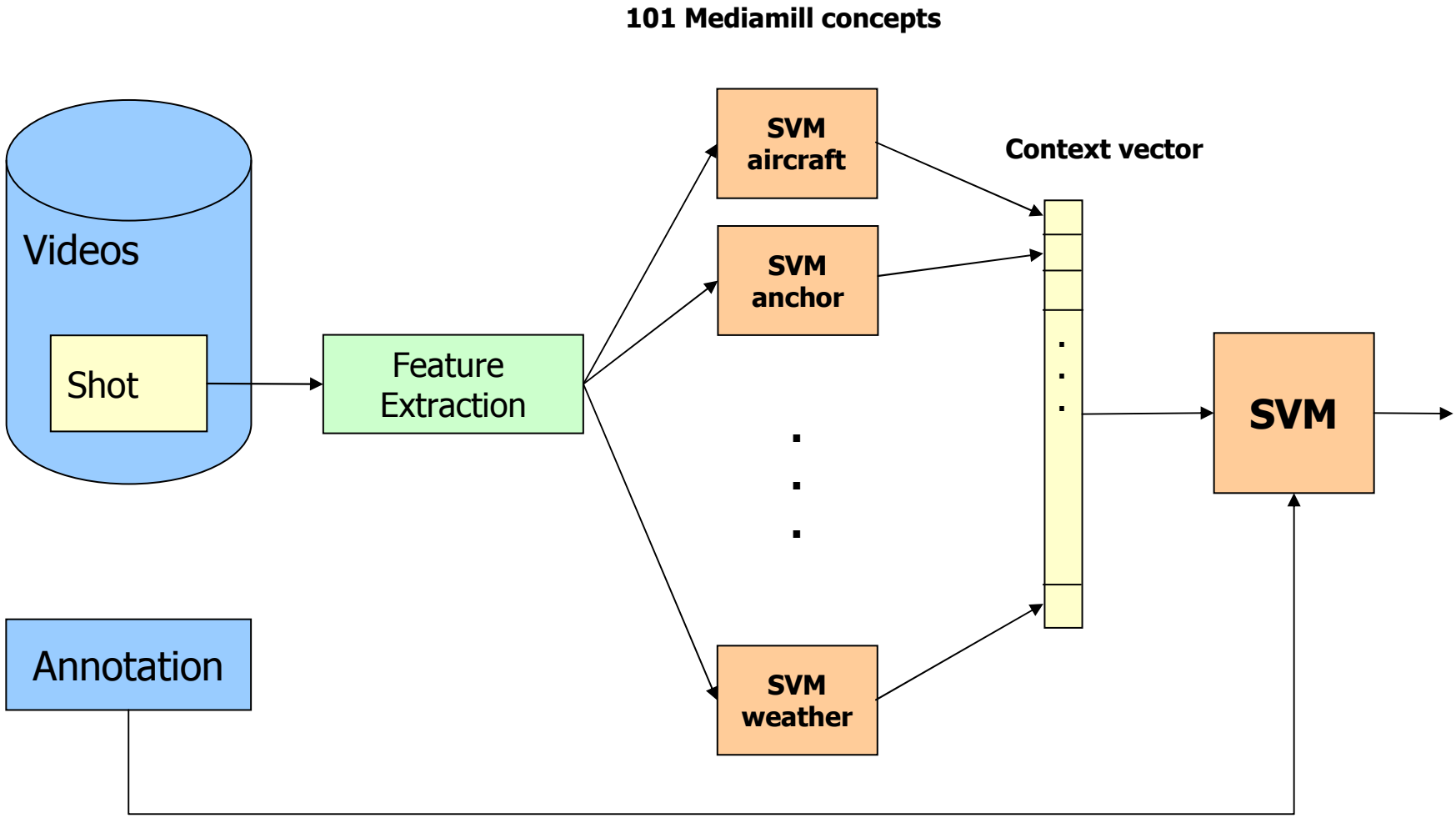
- # face sequences, length
- # frontal faces, # profile faces
- Shotsize

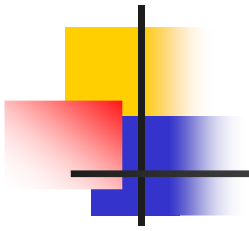
Audio Segmentation

- Hierarchical tree of support vector machines
  - Input:
    Audio low-level features (MFCCs, zero crossing rate, short time energy, spectrum flux, band periodicity, …)
  - Output:
    Audio type labels plus corresponding probabilities



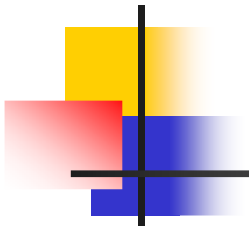
Context


101 Mediamill concepts
Videos
Shot
Feature Extraction
SVM aircraft
SVM anchor
SVM weather
Context vector
SVM
Annotation



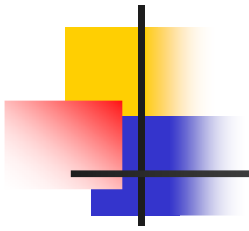
Outline

- Videana
- System overview
- Results
- Conclusions



Category a Runs

- a_ma1
  - Baseline system
  - TRECVID 2005 training set
  - Groundtruth from Mediamill challenge system
- a_ma5
  - Context features for 101 Mediamill concepts
  - Training set of Mediamill challenge system (subset of TRECVID 2005 training set)
- a_ma6
  - a_ma5 plus transductive learning

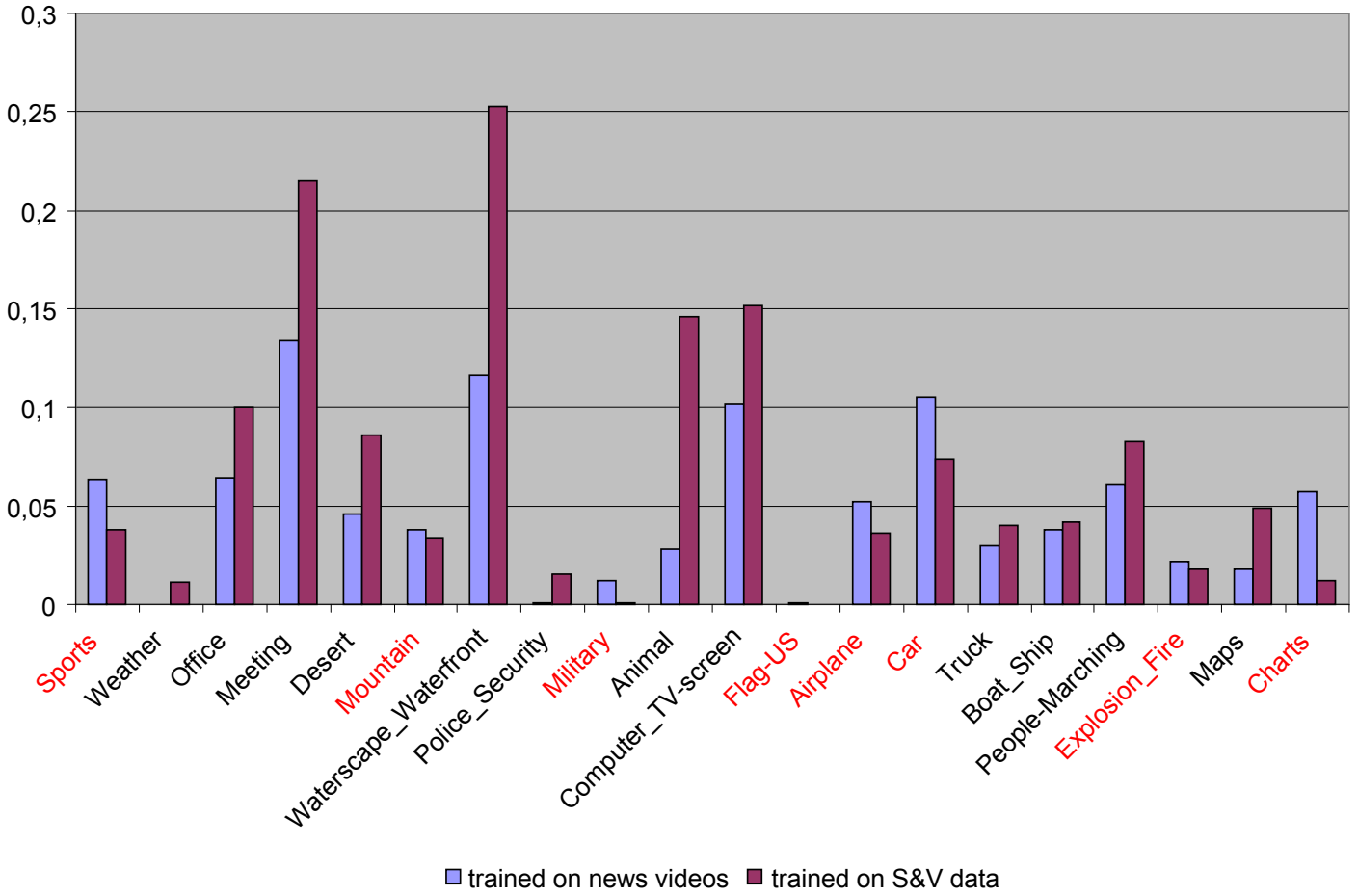


Category A Runs

- A_ma2
  - Baseline system
  - TRECVID 2007 training set
  - Merged annotations from active learning and MCQ-ECT-CAS
- A_ma3
  - A_ma2 plus additional distinction between color and gray-scale shots
- A_ma7 (additionally evaluated)
  - Context features for 101 Mediamill concepts
  - TRECVID 2005 and 2007 training set
- A_ma4
  - A_ma7 plus additional distinction between color and gray-scale shots

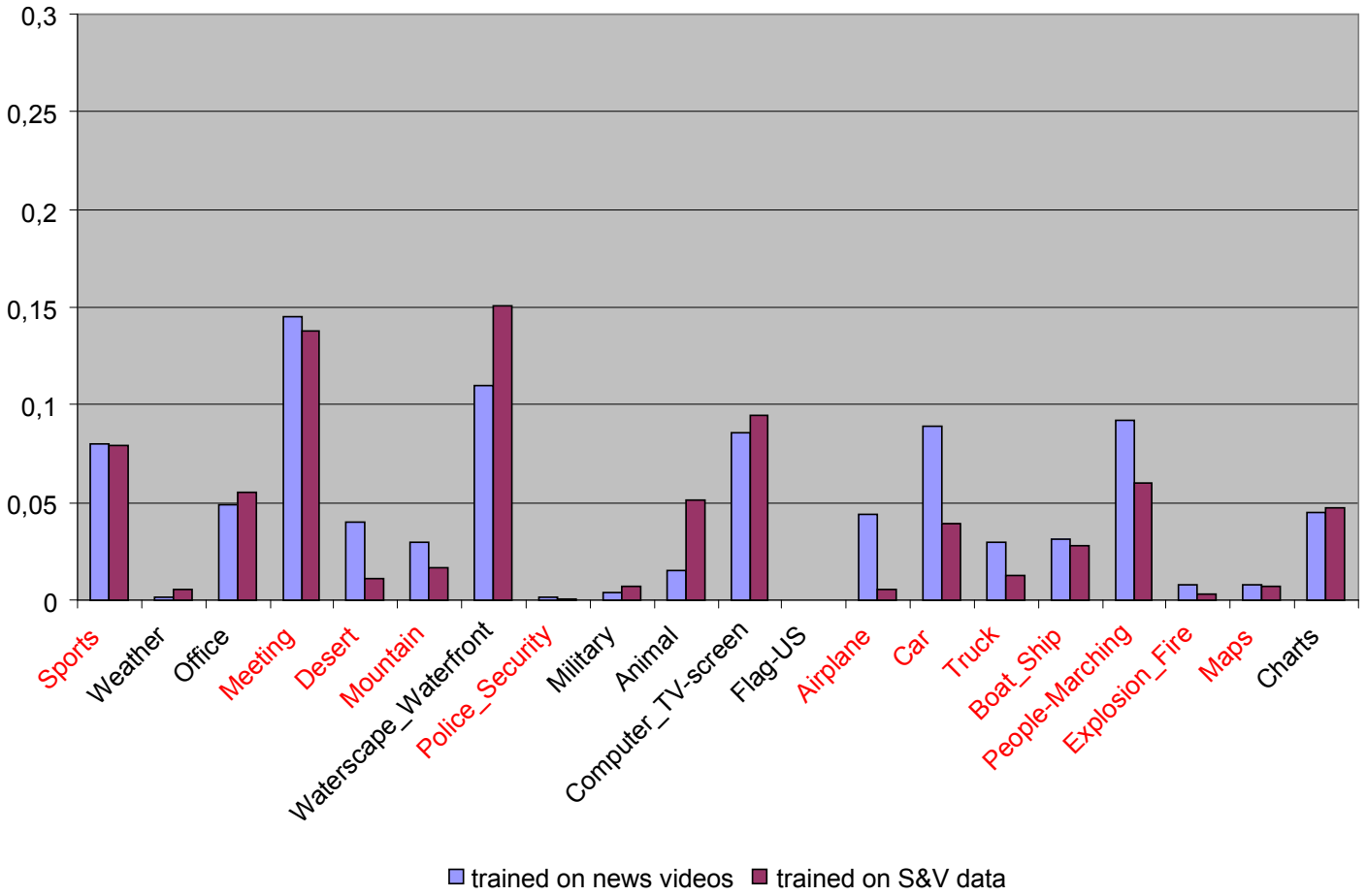
Baseline system

- Trained on news videos: 4.94% meaniAP
- Trained on S&V data: 7.03% meaniAP

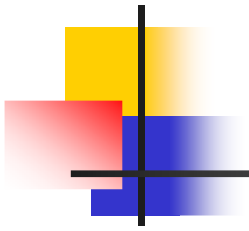


Context

- Trained on news videos: 4.55% meaniAP
- Trained on S&V data: 4.08% meaniAP



Well Generalizable Concept Models

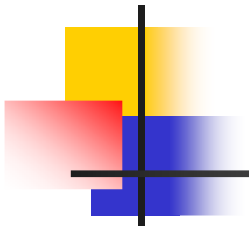


Baseline:

- Sports
- Mountain
- Military
- Flag-US
- Airplane
- Car
- Explosion-Fire
- Charts

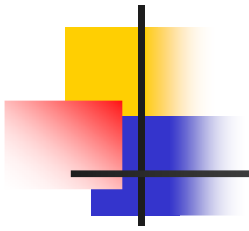
Context:

- Sports
- Meeting
- Desert
- Mountain
- Police_Security
- Airplane
- Car
- Truck
- Boat_Ship
- People-Marching
- Explosion-Fire
- Maps



Transductive Learning

- Appearance of several concepts strongly depends on video type
- Idea:
  Adapt the models based on news videos to the S&V data
- Using Transductive Support Vector Machines
  - Considers unlabeled test data in the training process
- No performance gain compared to the reference system (3.31% vs. 4.55%)
- Best results for “flag-us” and “airplane”

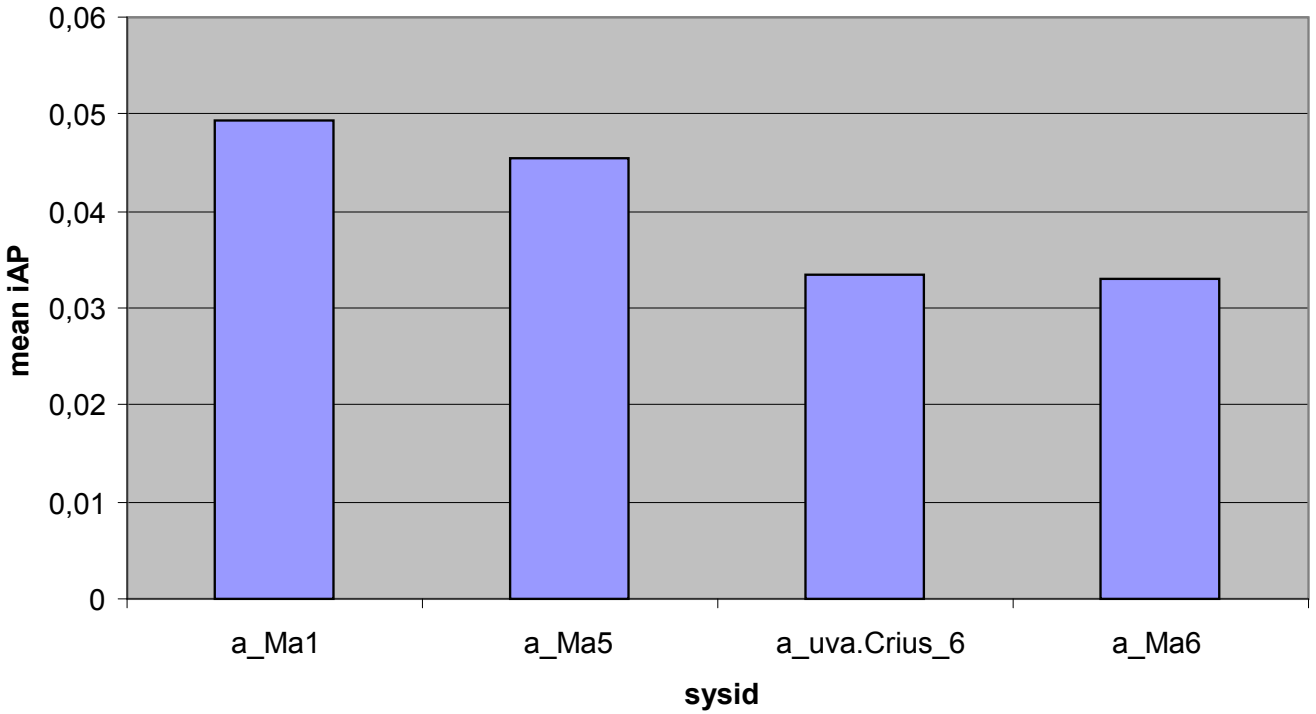


Distinguishing Color and Gray-Scale Images

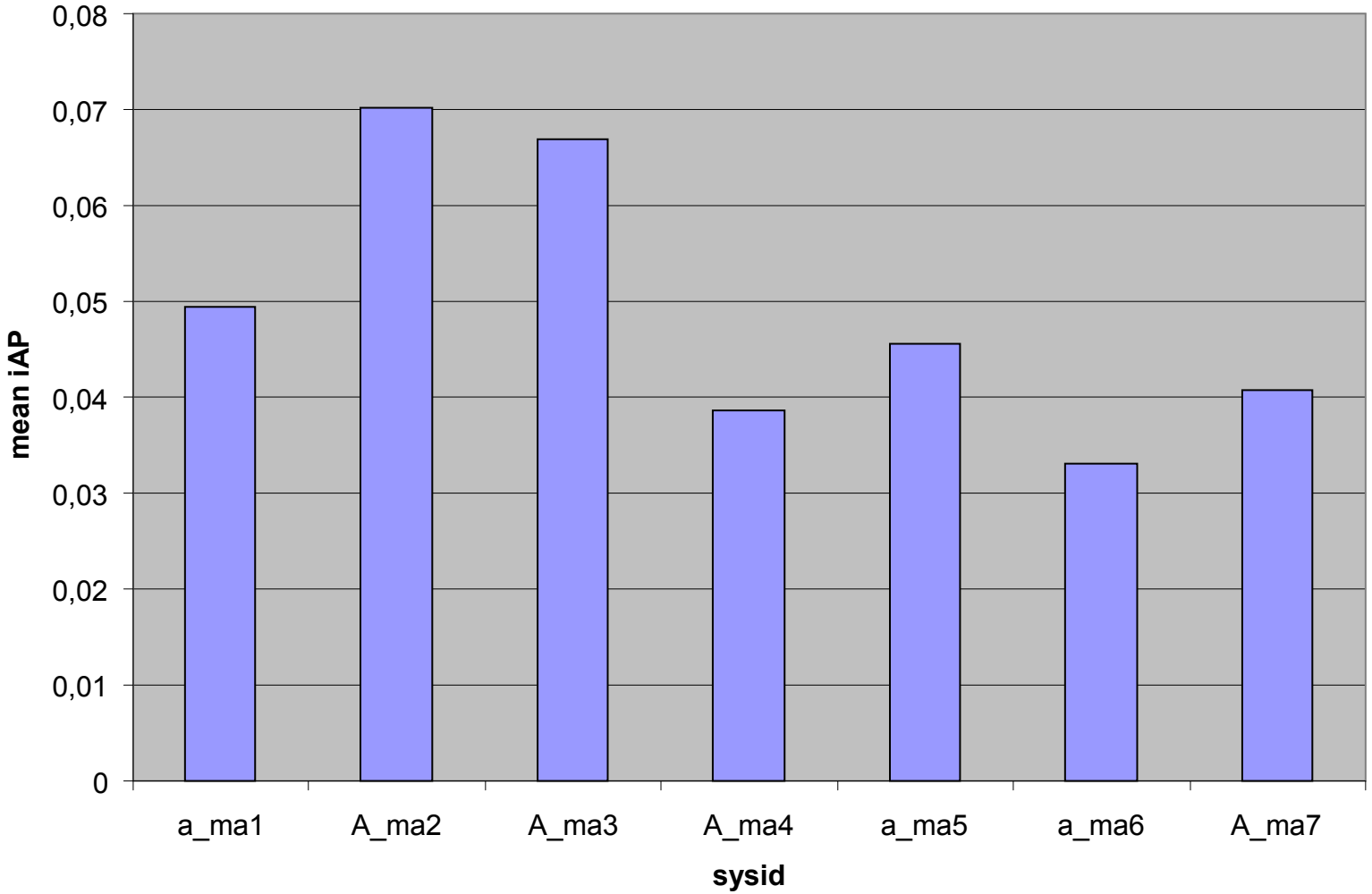
- Observation:
  - Many concepts strongly correlate with color features (Desert, Mountain, Road, Sky, …)
  - The S&V training set contains many gray-scale shots
- Using a support vector machine to classify keyframes
  - Accuracy: 94.9%
- Build separate models for both modalities
- Slight performance reduction for both systems
  - Baseline system A: 7.03% vs. 6.67%
  - Context A: 4.08% vs. 3.87%
- Similar performances for nearly all high-level features in comparison to the reference system

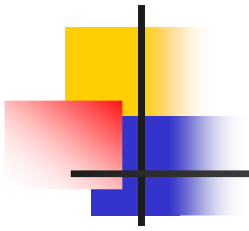
Systems not using S&V specific training data

Only 2 institutes submitted 4 runs



Submitted Runs





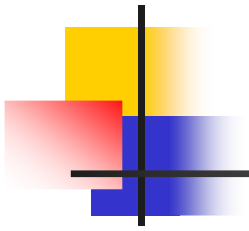
Summary of Results

- Our baseline system using the S&V training set achieved the clearly best result (7.03% meaniAP)
- High performance loss for context systems using the S&V training set
- The use of context concerning category “a” outperforms the corresponding category “A” system (4.55% vs. 4.08%)
- No improvement by transductive learning
- Systems distinguishing color and gray-scale images showed slight performance reductions

Number of Better Teams per Concept

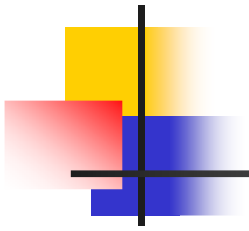
- Sports: 6
- Weather: 6
- Office: 13
- Meeting: 1
- Desert: 3
- Mountain: 16
- Waterscape_Waterfront: 10
- Police_Security: 5
- Military: 11
- Animal: 7
- Computer_TV-screen: 6
- Flag-US: 2
- Airplane: 4
- Car: 14
- Truck: 9
- Boat_Ship: 22
- People-Marching: 1
- Explosion_Fire: 6
- Maps: 13
- Charts: 8

➔ Second best result for Meeting and People-Marching



Outline

- Videana
- System overview
- Results
- Conclusions



Conclusions

- Limited generalization capabilities of systems trained on broadcast news videos to the S&V data
- However, several high-level feature models generalize very well to the S&V data (e.g. “car”, “explosion_fire”, “airplane”, “mountain” and “sports”)
- Our best system using the S&V training set achieved 7.03% meaniAP

Thanks for your attention!