Coloring Visual Codebooks for Concept Detection in Video

Koen van de Sande
Cees Snoek
Jan van Gemert
Jasper Uijlings
Jan-Mark Geusebroek
Theo Gevers
Arnold Smeulders

University of Amsterdam

MediaMill

VIDIVIDEO

Introduction

Concept detection:

- Machine learning based on image descriptors only

In a real-world video:

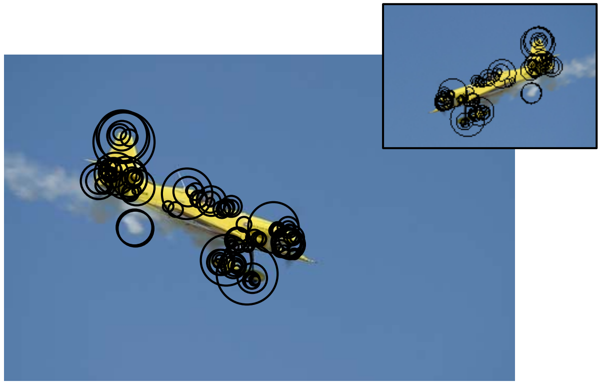
- Large variations in viewing and lighting conditions
  ➔ image description complicated

How do changes in viewpoint and illumination conditions affect concept detection?

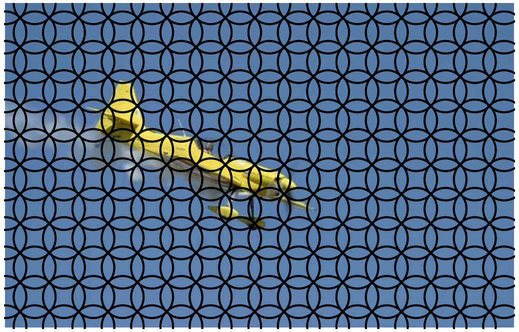
Lowe IJCV 2004
Mikolajczyk IJCV 2005
Zhang IJCV 2007
Marszalek VOC 2007

Viewpoint Changes

- Orientation and scale of object changes
- Salient point methods robustly detect regions



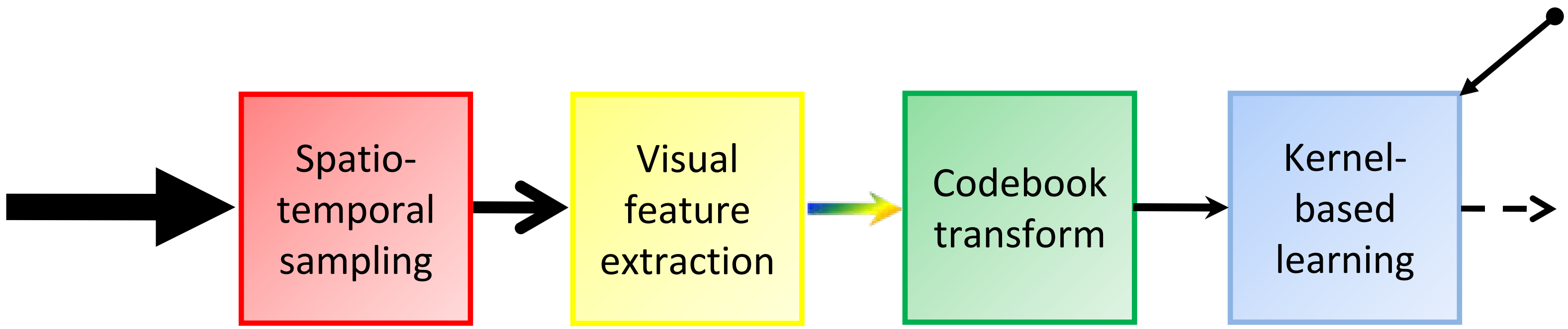
Harris-Laplace



Dense sampling

- INRIA-LEAR (VOC 2007 winner): preferred for concept detection accuracy are
  - Harris-Laplace salient points
  - Dense sampling

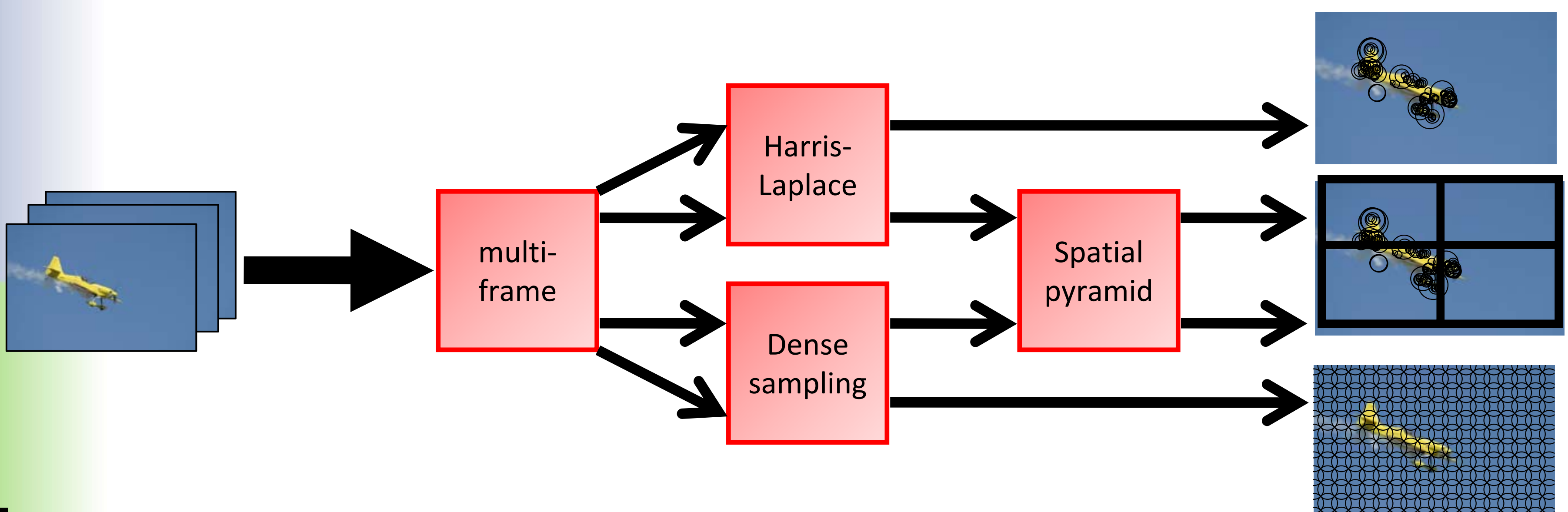
Concept Detection Stages



Spatio-Temporal Sampling

- Spatial pyramid
  - 1x1 whole image
  - 2x2 image quarters
  - 1x3 horizontal bars

- Temporal analysis of up to 5 frames per shot



Illumination Changes

van de Weijer PAMI 2006
Bosch CIVR 2007
Burghouts CVIU 2008
van de Sande CVPR 2008

Concept detection suffers from unstable region description

SIFT descriptor:

- Most well-known
- State-of-the-art performance
- Intensity-based descriptor: **no color**

Proposed color descriptors:

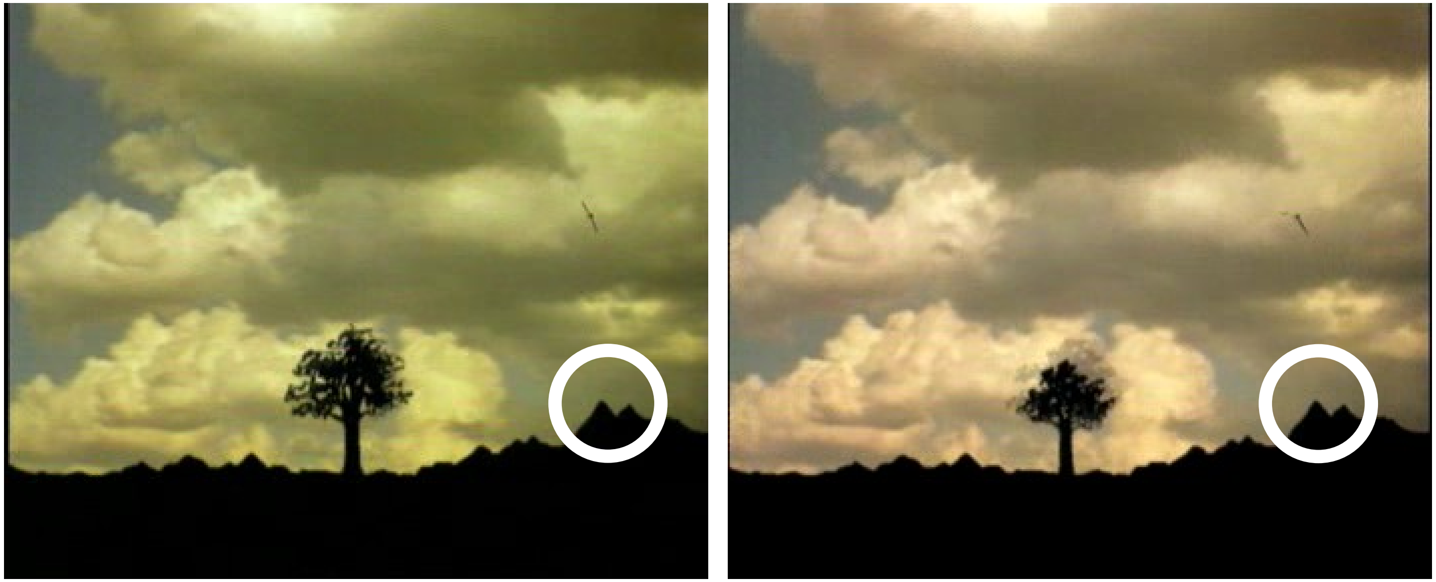
- HueSIFT, HSV-SIFT, OpponentSIFT, C-SIFT, *rg*SIFT
- Increase discriminative power
- Increase illumination invariance

Research questions

- What are the properties of these color descriptors?
- How do they perform?
- See the evaluation in our CVPR 2008 paper

Example: light color change

Transformed color SIFT descriptor is invariant



Von Kries 1970
Finlayson ICIP 2005

Invariance properties: Diagonal model

Lambertian reflectance model

$$\mathbf{f}(\mathbf{x}) = \int_{\omega} e(\lambda)\rho_k(\lambda)s(\mathbf{x},\lambda)d\lambda + \int_{\omega} a(\lambda)\rho_k(\lambda)$$

Corresponds to diagonal-offset model of illumination change

$$\boxed{\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix}} = \boxed{\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}} \boxed{\begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix}} + \boxed{\begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix}}$$

Illuminant parameters

Canonical illuminant

Unknown illuminant

Unified framework for modeling:

- Shadows
- Shading
- Light color changes
- Highlights
- Scattering

Color Descriptor Taxonomy

Invariance properties of the descriptors used

Descriptor	Light intensity change	Light intensity shift	Light intensity change and shift	Light color change	Light color change and shift
SIFT	+	+	+	+	+
OpponentSIFT	+/-	+	+/-	+/-	+/-
C-SIFT	+	+	+	+/-	+/-
*rg*SIFT	+	+	+	+/-	+/-
Transformed color SIFT	+	+	+	+	+

Invariant Visual Descriptors

Color SIFT:

- Intensity-based SIFT
- OpponentSIFT
- C-SIFT
- *rg*SIFT
- Transformed color SIFT

Add color, but also keep intensity information

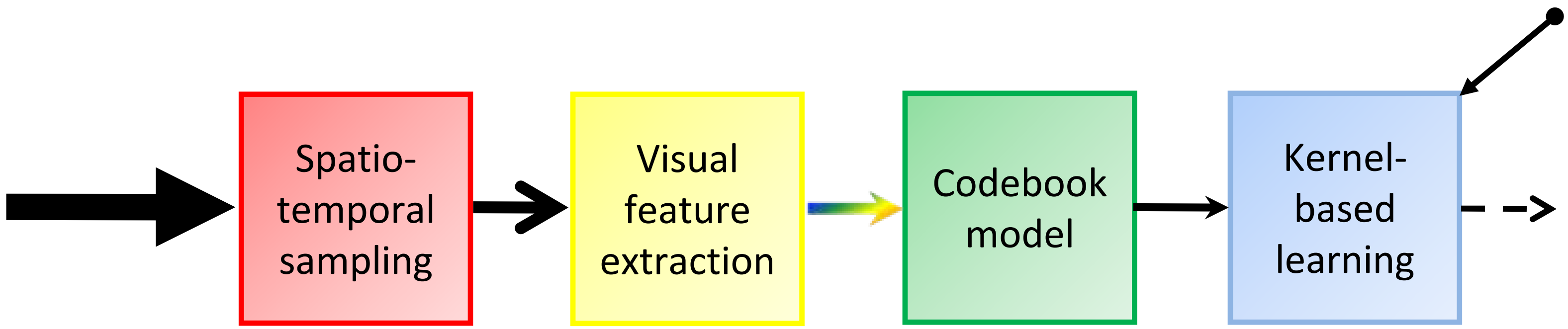
Visual Descriptors	MAP on TV2007test
Intensity SIFT	0,144
5x Color SIFT	0,155

relative +8%

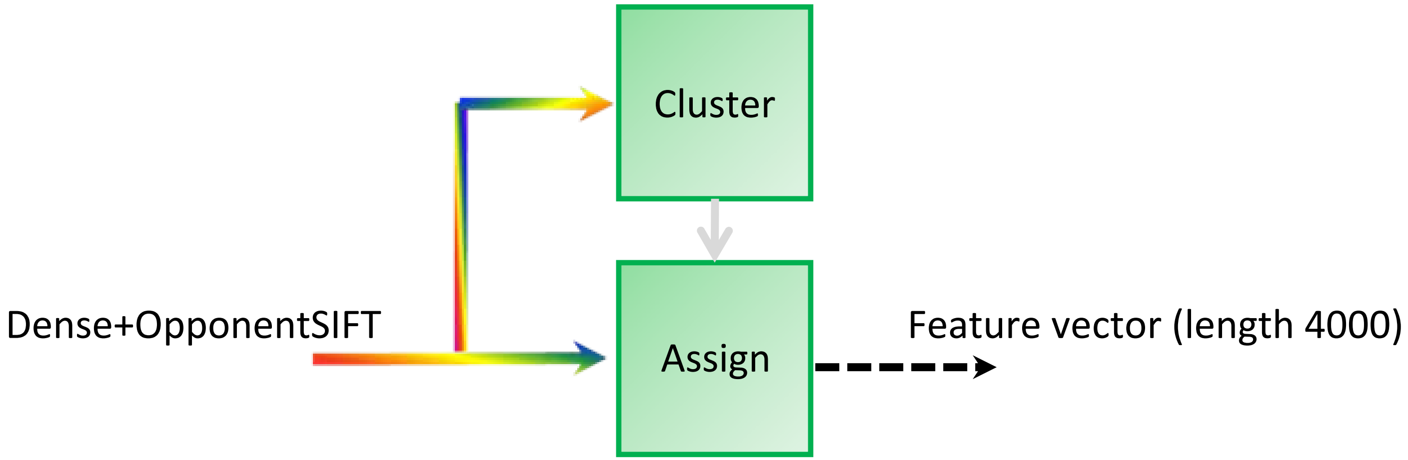
TV2007test results:

- Trained on TRECVID2007 development set
- Evaluated on TRECVID2007 test set
- TRECVID2007 development + test = 2008 development

Concept Detection Stages



Visual Codebook Model

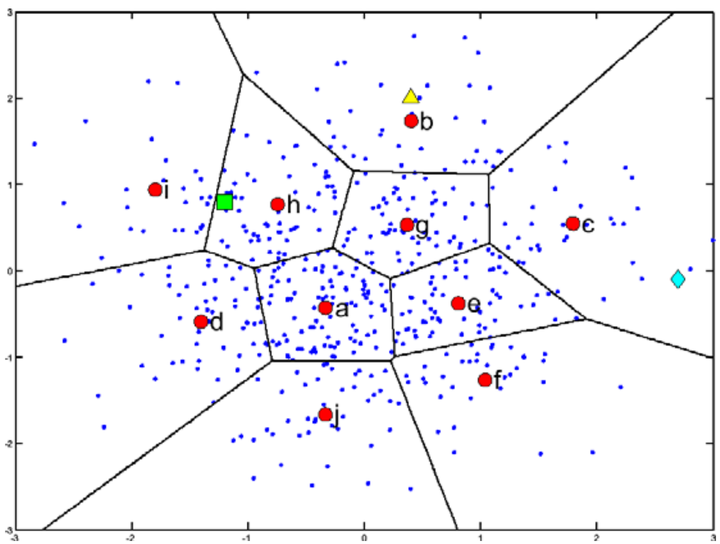


- Codebook consists of codewords
- Constructed with k-means clustering on descriptors
- We use 4,000 codewords per codebook

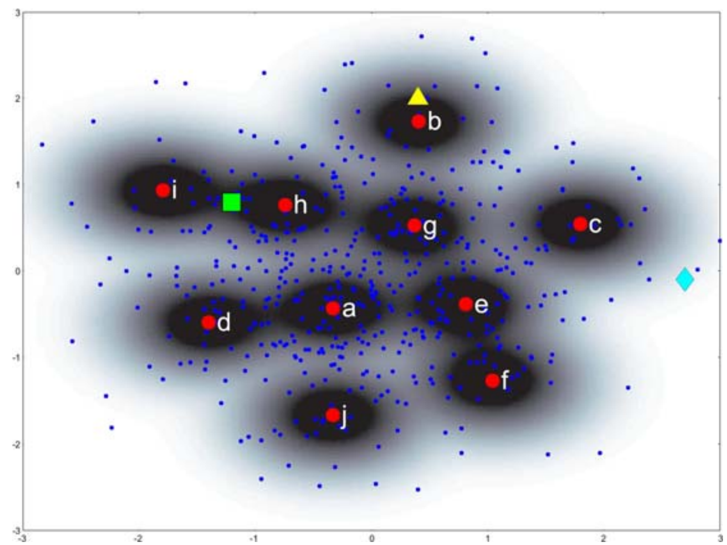
● Codeword

Codebook Assignment

Soft assignment using Gaussian kernel



Hard assignment

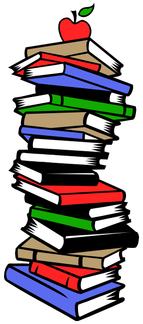


Soft assignment

Assignment	MAP on TV2007test
Hard	0,155
Soft	0,166

relative +7%

Codebook Library



Codebook	Sampling method	Descriptor	Construction	Assignment
#1	Dense	OpponentSIFT	K-means	Soft
#2	Harris-Laplace	SIFT	Radius-based	Soft
#3	Dense	*rg*SIFT	K-means	Hard
…	Dense	C-SIFT	K-means	Hard

Single codebook depends on

- Sampling method
- Descriptor
- Codebook construction method
- Codebook assignment

Codebook library is…

- a configuration of several codebooks

Codebook Library (cont'd)

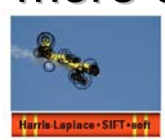
For a frame:

- Each codebook in the library has feature vector of length 4,000

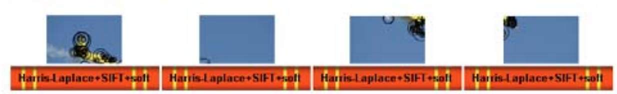
- Final feature vector is concatenation (4 books ~ length 16,000)

- Spatial pyramid adds more dimensions:

•1x1 4,000



•2x2 16,000

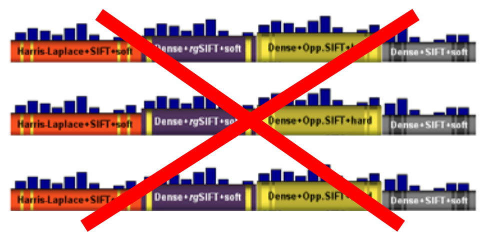


•1x3 12,000

- Feature vector length easily >100,000…

SVM kernel trick: precompute kernel

SVM learning does not need feature vectors



SVM learning needs distance between vectors only:

K( , ) = $e^{-\gamma\, dist(}$  ,  $^{)}$

Very large decrease in computation time

- Precompute the SVM kernel matrix
- Long vectors possible: only need 2 in memory at once
- Parameter optimization re-uses precomputed matrix

Impact of annotations

Ours = common annotation effort + ICT-CAS + verifying positives

Codebook library	Ours* (type B)	Common ann. effort* (type A)
3x Color SIFT	0,152	0,152
5x Color SIFT	0,155	0,155

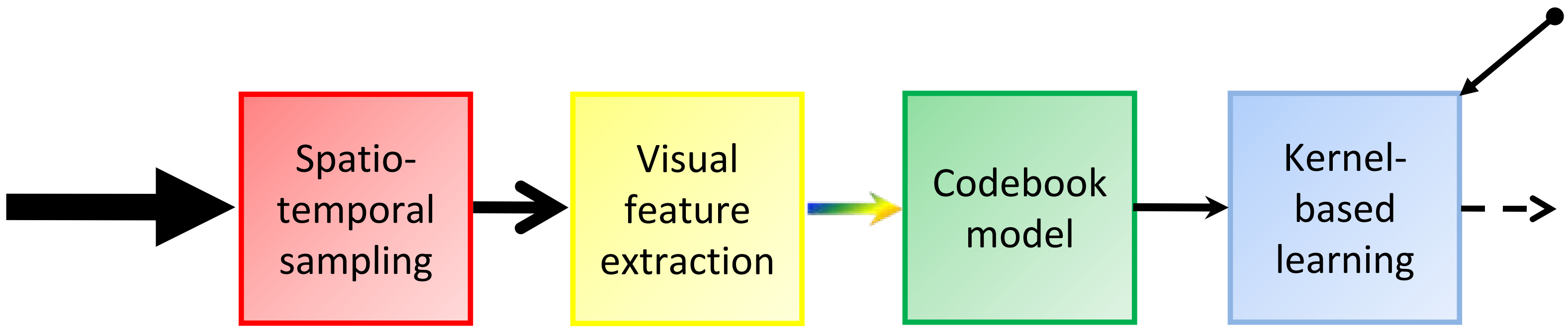
*MiAP on TV2008test

Add a digit…

Codebook library	Ours*	Common ann. effort*
3x Color SIFT	0,1516	0,1521
5x Color SIFT	0,1548	0,1549

On average, didn't help

Concept Detection Stages



Robust Temporal Approach

- No cloud computing yet: need to be efficient ☺
- Process 5 frames per shot in test set
- Linear increase in computation: x5

Codebook library	Frames/shot	MiAP on TV2008test
3x Color SIFT	1	0,152
3x Color SIFT	5	0,184

relative +20%

- In 2005 paper 7.5% to 38% improvement noted for multi-frame (worst-case vs. best-case using oracle)
- **Robust color SIFT *with* temporal = ~20% improvement**

The Good

- Close-up of hands

- Boats and ships

- Cityscape

The Bad

- Emergency Vehicle (only 46 examples, many at night)

- Bus (only 64 examples)

… and the trivial

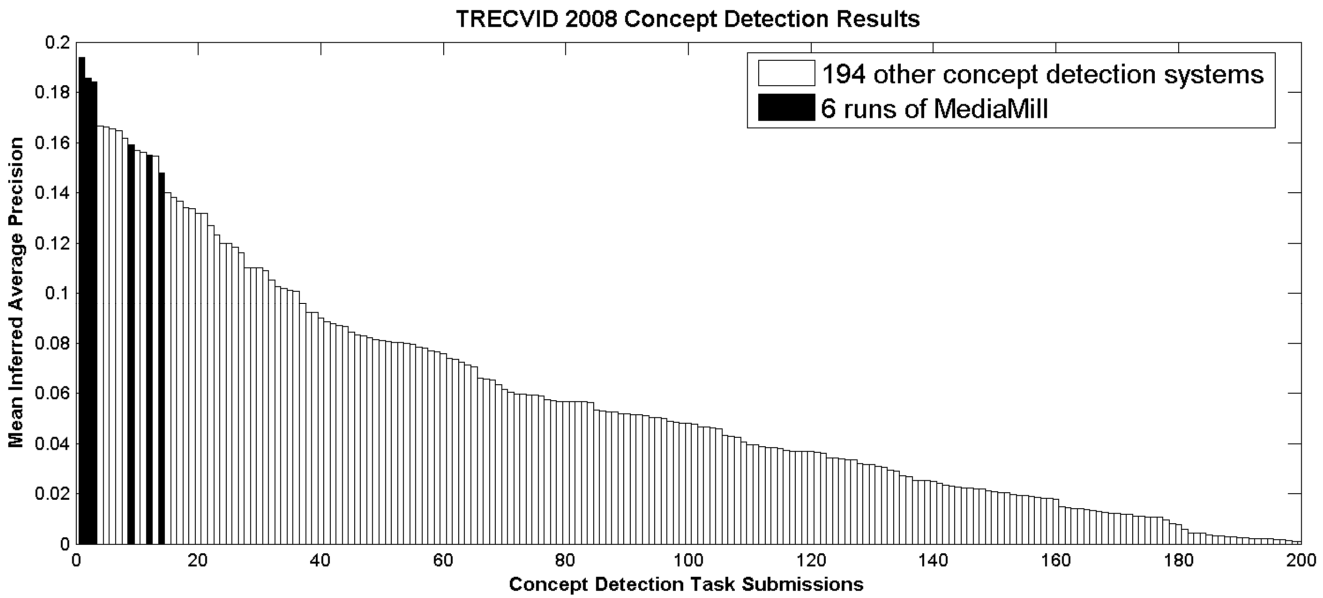
- Dog (in trailer)

- Flower (in trailer)

- Mountain (in trailer)

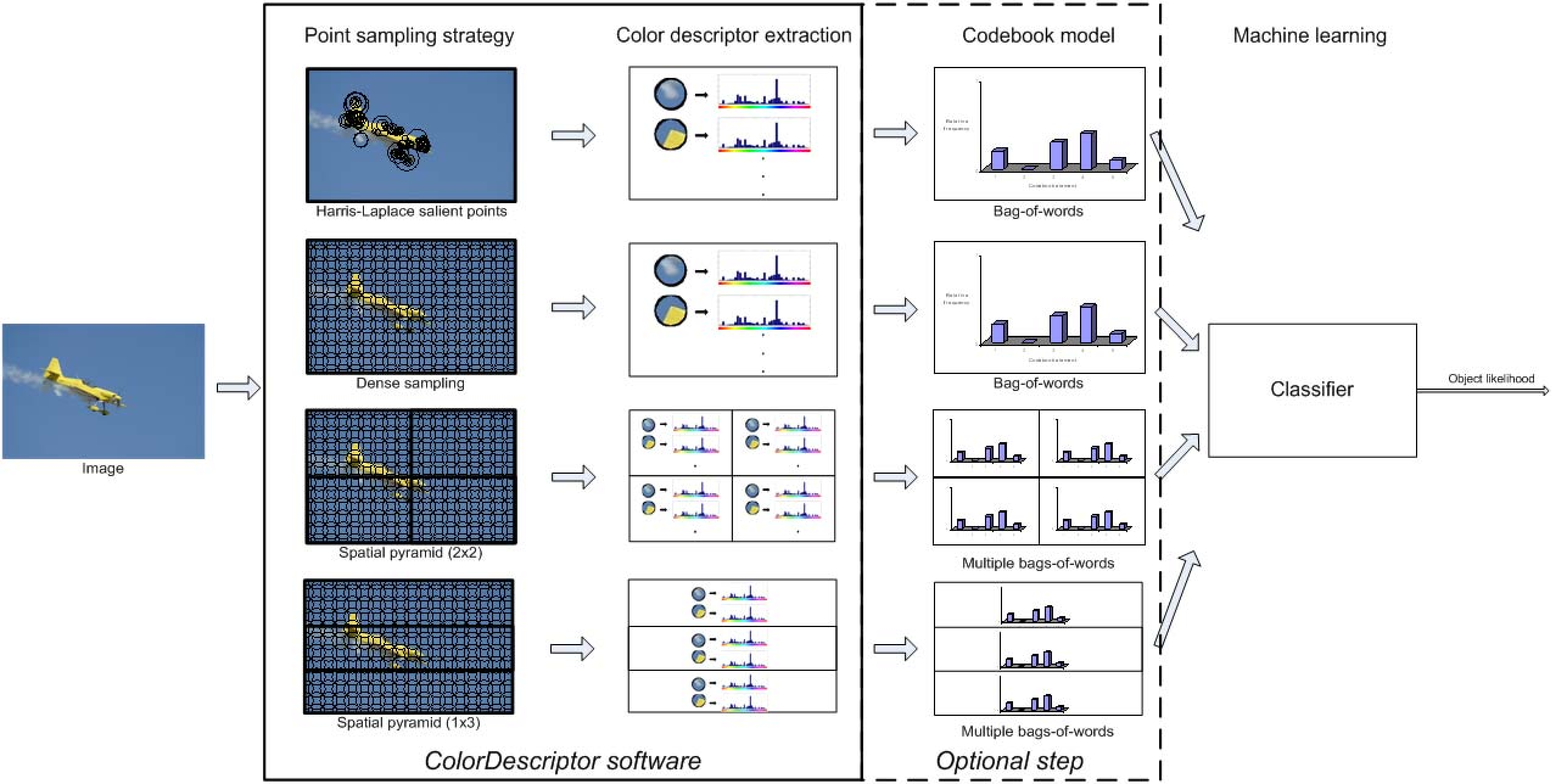
Conclusions

- Illumination conditions affect concept detection
- SIFT+colorSIFT improves ~8%
- Soft codebook assignment improves ~7%
- Robust colorSIFT with simple multi-frame improves ~20%:
  - Room for more advanced methods in TRECVID 2009
- Precomputed kernel matrix reduces SVM computation time
- Near-duplicates from trailers hamper progress:
  - We suggest to exclude them, or count only once



ColorDescriptor software

for object and scene categorization



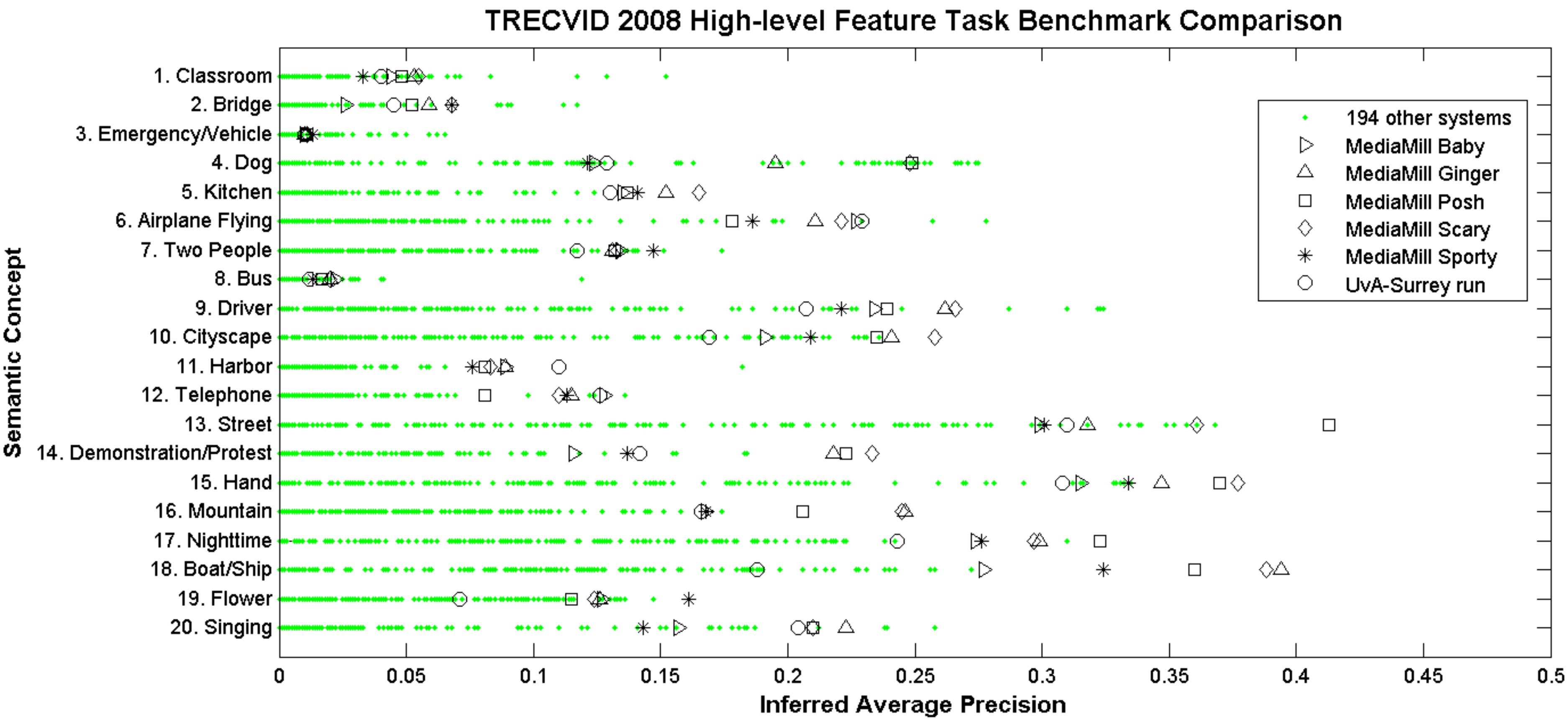
Created by Koen van de Sande

Visit http://www.science.uva.nl/~ksande/ for color descriptor software

References

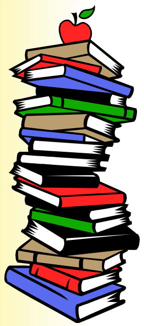
- K. E. A. van de Sande, T. Gevers and C. G. M. Snoek, "*Evaluation of Color Descriptors for Object and Scene Recognition*", CVPR 2008
- M. Marszalek, C. Schmid, H. Harzallah and J. van de Weijer, "*Learning Object Representations for Visual Object Class Recognition*", Visual Recognition Workshop in conjunction with ICCV 2007
- J.C. van Gemert, J.M. Geusebroek, C.J. Veenman, A.W.M. Smeulders, "*Kernel Codebooks for Scene Categorization*", ECCV 2008
- K. Mikolajczyk and C. Schmid, "*A Performance Evaluation of Local Descriptors*", PAMI 2005
- D. G. Lowe, "*Distinctive Image Features from Scale-Invariant Keypoints*", IJCV 2004
- J. Zhang, M. Marszalek, S. Lazebnik and C. Schmid, "*Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study*", IJCV 2007
- C. G. M. Snoek et al, "*The MediaMill TRECVID 2008 Semantic Video Search Engine*", TRECVID Workshop 2008

Results per Concept



Codebook Library Definitions

- Intensity SIFT

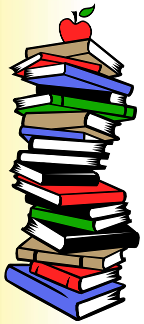


Codebook	Sampling method	Descriptor	Construction	Assignment
#1	Dense	SIFT	K-means	Hard
#2	Harris-Laplace	SIFT	K-means	Hard

- 5x Color SIFT / Hard

Codebook	Sampling method	Descriptor	Construction	Assignment
#1	Dense	OpponentSIFT	K-means	Hard
#2	Harris-Laplace	OpponentSIFT	K-means	Hard
#3	Dense	Transformed color SIFT	K-means	Hard
#4	Harris-Laplace	Transformed color SIFT	K-means	Hard
#5	Dense	SIFT	K-means	Hard
#6	Harris-Laplace	SIFT	K-means	Hard
#7	Dense	C-SIFT	K-means	Hard
#8	Harris-Laplace	C-SIFT	K-means	Hard
#9	Dense	*rg*SIFT	K-means	Hard
#10	Harris-Laplace	*rg*SIFT	K-means	Hard

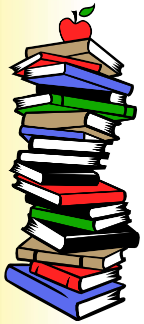
Codebook Library Definitions (2)



- 5x Color SIFT / Soft

Codebook	Sampling method	Descriptor	Construction	Assignment
#1	Dense	OpponentSIFT	K-means	Soft
#2	Harris-Laplace	OpponentSIFT	K-means	Soft
#3	Dense	Transformed color SIFT	K-means	Soft
#4	Harris-Laplace	Transformed color SIFT	K-means	Soft
#5	Dense	SIFT	K-means	Soft
#6	Harris-Laplace	SIFT	K-means	Soft
#7	Dense	C-SIFT	K-means	Soft
#8	Harris-Laplace	C-SIFT	K-means	Soft
#9	Dense	*rg*SIFT	K-means	Soft
#10	Harris-Laplace	*rg*SIFT	K-means	Soft

Codebook Library Definitions (3)



- 3x Color SIFT

Codebook	Sampling method	Descriptor	Construction	Assignment
#1	Dense	OpponentSIFT	K-means	Soft
#2	Harris-Laplace	OpponentSIFT	K-means	Soft
#3	Dense	Transformed color SIFT	K-means	Soft
#4	Harris-Laplace	Transformed color SIFT	K-means	Soft
#5	Dense	SIFT	K-means	Soft
#6	Harris-Laplace	SIFT	K-means	Soft

Positive Examples Needed

Concept	Relative #positive examples
TwoPeople	||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Street	|||||||||||||||||||||||||||||||||||||||||||||||||||||
Hand	||||||||||||||||||||||||||||||||||||||||||||
Flower	|||||||||||||||||
Singing	||||||||||||||||
BoatShip	|||||||||||||
Driver	|||||||||||||
Nighttime	|||||||||||||
Mountain	||||||||
Harbor	|||||||
Classroom	|||||||
Telephone	|||||||
DemonstrationOrProtest	|||||
Cityscape	||||
Bridge	||||
Kitchen	||||
Dog	|||
EmergencyVehicle		
AirplaneFlying		
Bus		

= highest overall infAP for MediaMill

Annotation effects: 5x Color SIFT

	Type B (ours)	Type A (common ann. effort)
Classroom	0,044	0,035
Bridge	0,026	0,049
EmergencyVehicle	0,010	0,016
Dog	0,124	0,128
Kitchen	**0,135**	0,109
AirplaneFlying	**0,227**	0,181
TwoPeople	0,134	0,128
Bus	0,022	0,014
Driver	0,234	**0,276**
Cityscape	0,191	0,195
Harbor	0,089	0,094
Telephone	0,128	0,149
Street	0,299	0,295
DemonstrationOrProtest	0,116	0,100
Hand	**0,315**	0,286
Mountain	0,168	**0,249**
Nighttime	**0,274**	0,232
BoatShip	0,277	0,273
Flower	0,127	**0,155**
Singing	0,157	0,134
MiAP	**0,1548**	**0,1549**

Annotation effects: 3x Color SIFT

	Type B (ours)	Type A (common ann. effort)
Classroom	0,044	0,044
Bridge	0,053	0,076
EmergencyVehicle	0,010	0,010
Dog	0,122	0,123
Kitchen	**0,132**	0,115
AirplaneFlying	**0,212**	0,154
TwoPeople	0,128	0,128
Bus	0,016	0,009
Driver	0,209	**0,258**
Cityscape	0,210	0,216
Harbor	0,064	0,059
Telephone	0,109	0,124
Street	0,276	0,269
DemonstrationOrProtest	0,138	0,145
Hand	0,279	0,271
Mountain	0,162	**0,224**
Nighttime	**0,282**	0,244
BoatShip	0,308	0,309
Flower	0,104	0,089
Singing	0,177	0,178
MAP	**0,1516**	**0,1521**