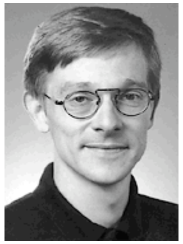
Models for Joint Labeling of Objects and Scenes

TECHNISCHE UNIVERSITÄT DARMSTADT



Bernt Schiele

Department of Computer Science
TU Darmstadt

thanks to my collaborators

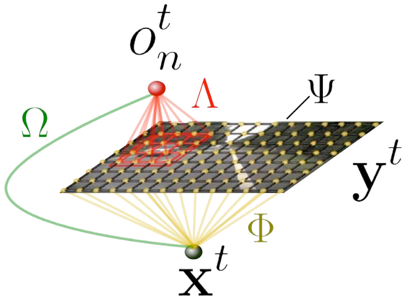
Julia Vogel

Christian Wojek

Overview

- Semantic Scene Modeling [ijcv07,tap'06]
  - natural **scene categorization is not enough**
  - aim for typicality ranking instead !
  - joint work with **Julia Vogel**

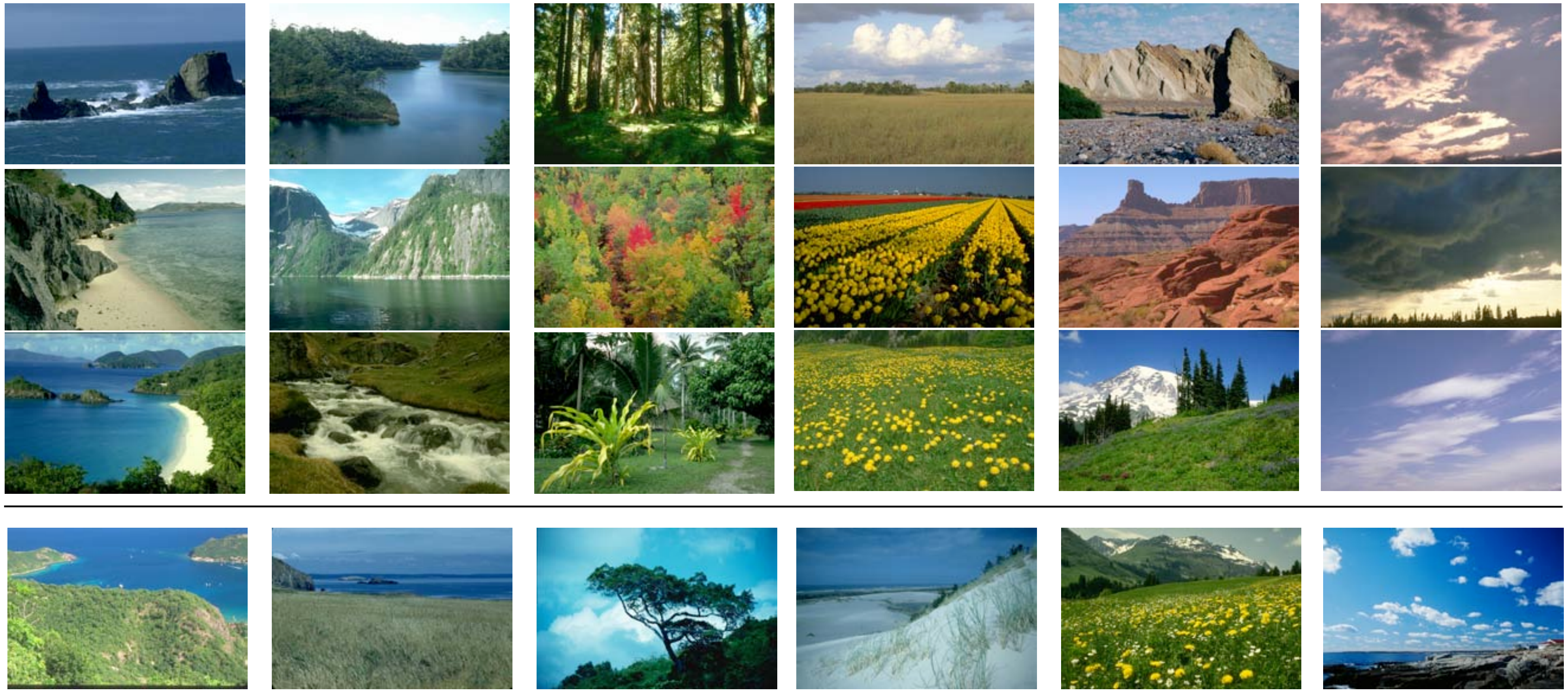
- Joint Labeling of Objects and Scenes [eccv08]
  - dynamical conditional random field model
  - joint work with **Christian Wojek**



Natural Scene Modeling

Photographic Images
- indoors
- outdoors
  - artificial
  - natural
    - coasts
    - rivers/lakes
    - forests
    - plains
    - mountains
    - sky/clouds

Basic Level Scene Categories



Typical

Less typical

Semantic Modeling

Local Semantic Concepts*

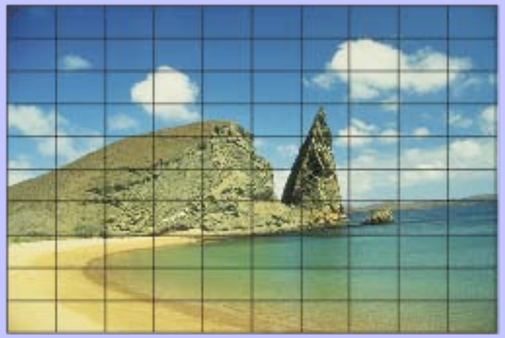
*inspired by [Mojsilovic et al., 2004]

sky water grass trunks foliage field rocks flowers sand

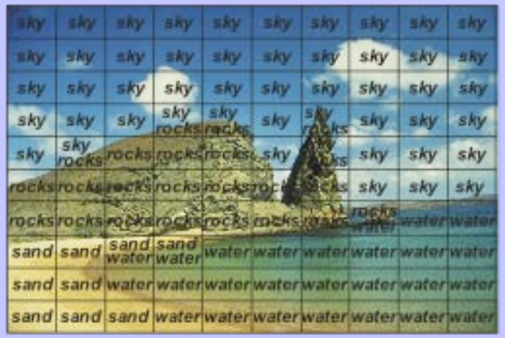
Global image representation, e.g. for categorization or ranking

Database Images

10x10 Grid



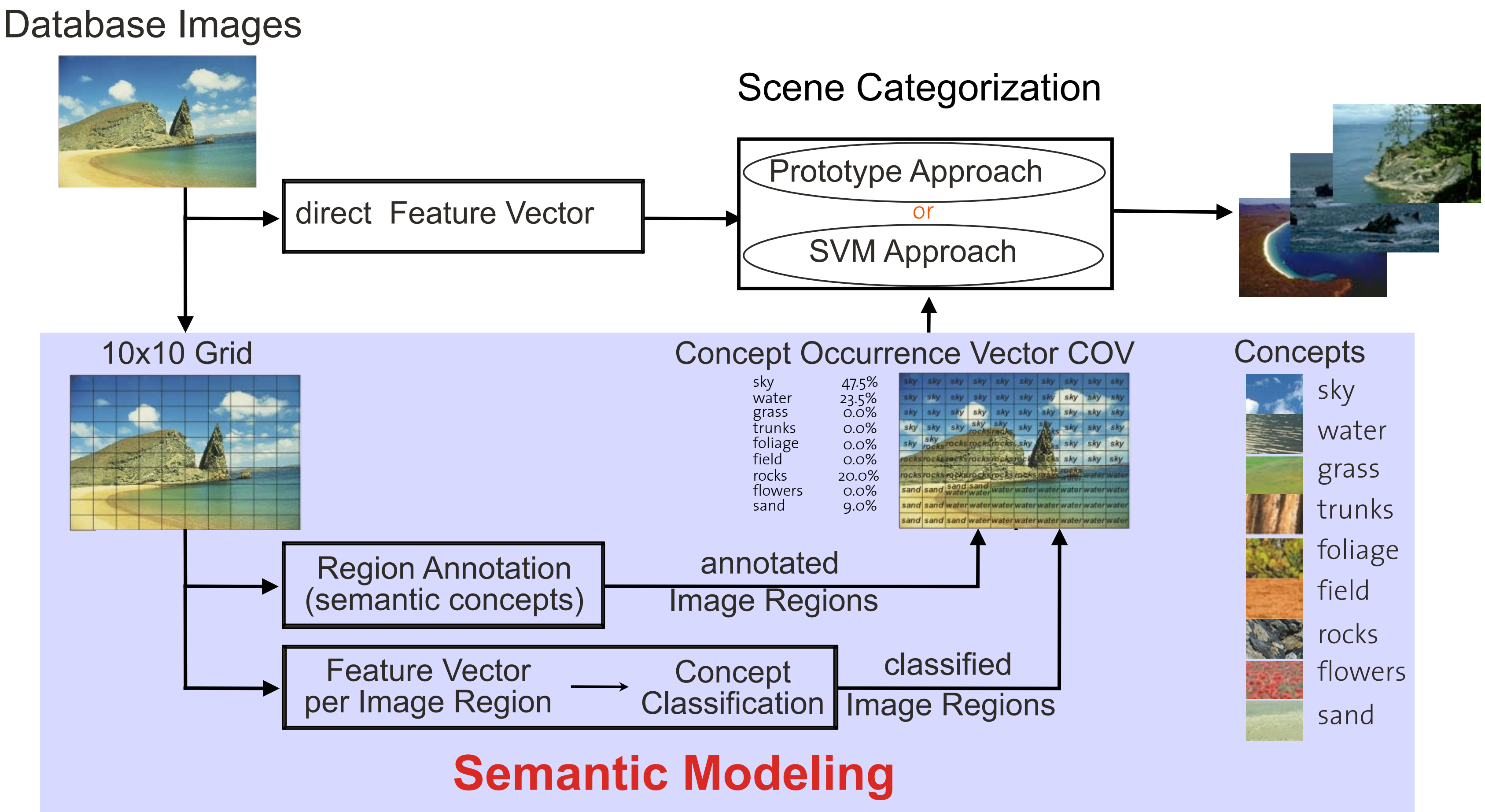
Semantic Labeling



Concept Occurrence Vector

sky	47.5%
water	23.5%
grass	0.0%
trunks	0.0%
foliage	0.0%
field	0.0%
rocks	20.0%
flowers	0.0%
sand	9.0%

Categorization Experiments

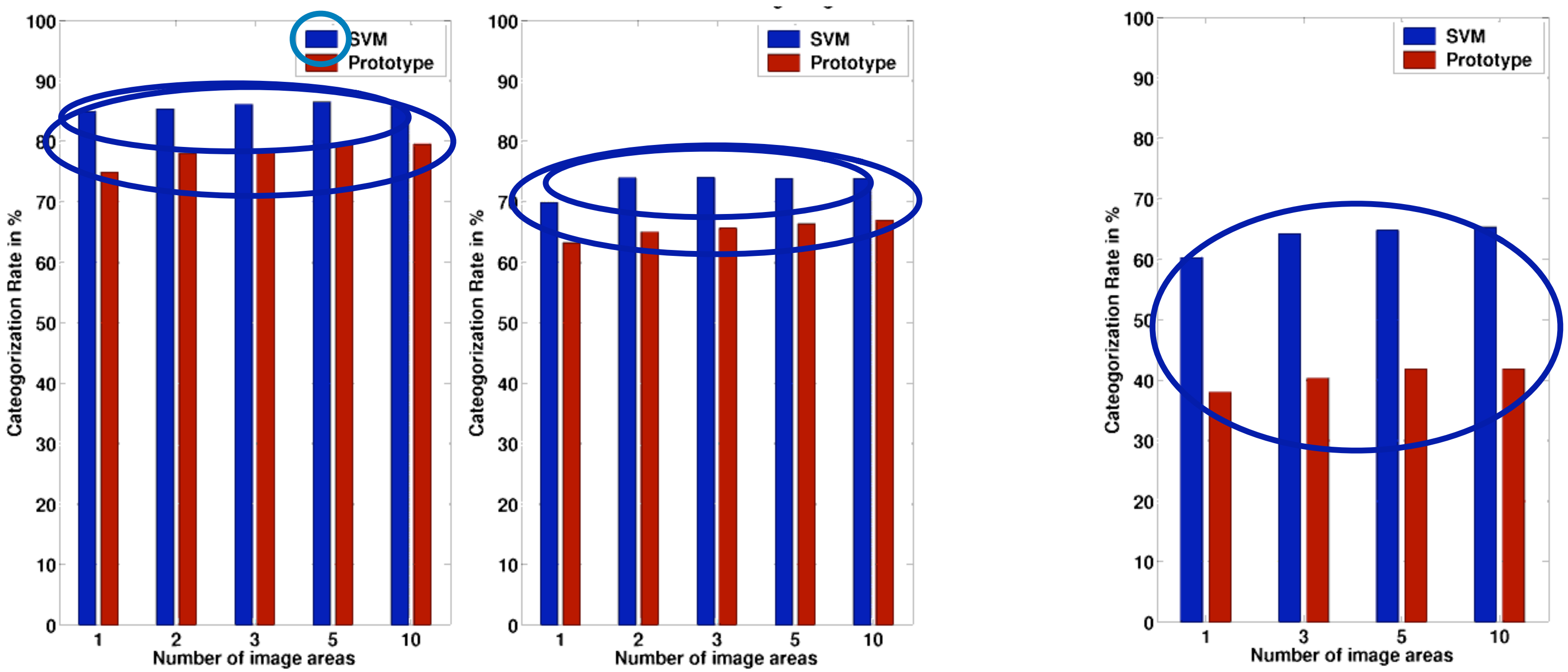


1. Semantic Modeling vs. Direct Feature Extraction
2. Annotated vs. Classified Image Regions
3. Prototype vs. SVM Classifier

Categorization Results

Semantic Modeling | No Semantic Modeling

Annotated Regions | Classified Regions



1. **Support-Vector Machines outperform Prototypes.**
2. **Semantic Modeling improves results considerably.**
3. **Fully automatic categorization at 74% categorization rate**

But: Benchmark (annotated regions) at only 86.4% categorization rate.

Semantic Analysis

Benchmark at only 86.4% categorization rate

- Classification problem? Inherent problem?
- Analyze semantically!

Three points for semantic analysis:

✓ 1. Visual inspection of mis-categorizations

"Correct" category in parentheses

forests
(rivers/lakes)

rivers/lakes
(mountains)

plains
(coasts)

mountains
(plains)

Semantic Analysis

Benchmark at only 86.4% categorization rate

- Classification problem? Inherent problem?
- Analyze semantically!

Three points for semantic analysis:

✓ 1. Visual inspection of mis-categorizations

✓ 2. Confusions of benchmark: Make sense semantically?

	coasts	rivers	forests	mount	plains	sky
coasts	**80.3**	14.1	0.7	3.5	0.7	0.7
rivers/lakes	18.0	**73.0**	3.6	0.9	3.6	0.9
forests	0.0	1.9	**95.1**	1.9	1.0	0.0
mountains	0.8	0.0	0.8	**91.6**	5.3	1.5
plains	0.6	2.2	0.6	6.7	**89.4**	0.6
sky/clouds	0.0	0.0	0.0	5.9	0.0	**94.1**

Confusion matrix

Semantic Analysis

Benchmark at only 86.4% categorization rate

- Classification problem? Inherent problem?
- Analyze semantically!

Three points for semantic analysis:

✓ 1. Visual inspection of mis-categorizations

✓ 2. Confusions of benchmark: Make sense semantically?

✓ 3. Rank Statistics: Rankings meaningful?

	coasts	rivers	forests	mount	plains	sky
coasts	**80.3**	14.1	0.7	3.5	0.7	0.7
rivers/lakes	18.0	**73.0**	3.6	0.9	3.6	0.9
forests	0.0	1.9	**95.1**	1.9	1.0	0.0
mountains	0.8	0.0	0.8	**91.6**	5.3	1.5
plains	0.6	2.2	0.6	6.7	**89.4**	0.6
sky/clouds	0.0	0.0	0.0	5.9	0.0	**94.1**

Confusion matrix

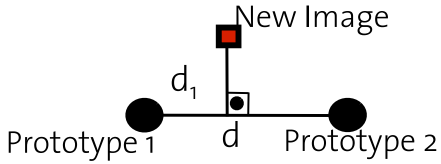
1	2	3	4	5	6
80.3	97.1	99.3	99.3	100.0	100.0
73.0	95.5	96.4	99.1	100.0	100.0
95.1	98.1	99.0	100.0	100.0	100.0
91.6	98.5	98.5	100.0	100.0	100.0
89.4	98.3	98.9	100.0	100.0	100.0
94.1	100.0	100.0	100.0	100.0	100.0
86.4	97.7	98.6	99.7	100.0	100.0

Rank Statistics

Typicality Transitions

Use normalized Euclidean distance D between two categories for ranking.

$$D = \frac{d_1}{d}$$



How do humans rank these images?

forests ⇨ mountains


D=0.05

D=0.11

D=0.48

D=0.62

D=0.87

mountains ⇨ rivers/lakes


D=0.11

D=0.40

D=0.67

D=0.77

D=0.82

rivers/lakes ⇨ forests


D=0.06

D=0.29

D=0.34

D=0.81

D=0.95

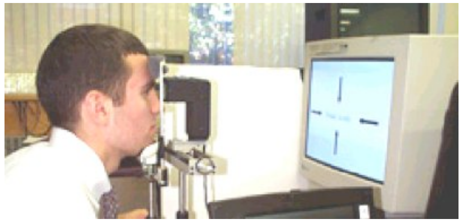
Psychophysical Experiments

Experiments in collaboration with Schwaninger/Hofer, University of Zurich

How do humans perceive natural scenes?

Setup:

- Dimly lit room, chin rest
- 250 images: coasts, rivers/lakes, forests, plains, mountains



Experiment 1: Categorization

- Assign image as quickly as possible to one of the five categories.
- 20 participants

Experiment 2: Typicality Rating

- How typical is image relative to each of the categories?
- 10 participants

Results of Human Studies

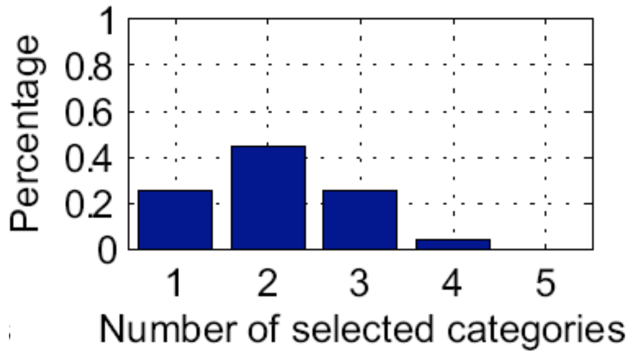
1. Participants very consistent in their decisions (Cronbach's $\alpha > 0.9$)
2. Typicality ranking consistent over participants (Spearman's rank correlation $r_s > 0.6$)

	Study 1	Study 2	
	Cronbach's α	Cronbach's α	Rank Correlation r_s
coasts	0.98	0.98	0.69
rivers/lakes	0.97	0.98	0.78
forests	0.99	0.97	0.81
plains	0.99	0.97	0.68
mountains	0.98	0.94	0.65

Inter-rater reliabilities

3. Many images are (at least partially) semantically ambiguous !

Response Distribution
Study 1: Categorization



Results of Human Studies (2)

Unanimously


rivers/lakes


plains


mountains

"Fifty,fifty."


45% forests
55% plains


45% plains
55% mountains


60% coasts
40% rivers/lakes

Distributed over three categories


25% forests
40% plains
35% mountains


10% rivers/lakes
55% forests
35% mountains


75% rivers/lakes
10% coasts
15% mountains

Conclusion: Aim for automatic typicality ranking.

Automatic Typicality Ranking: PPD

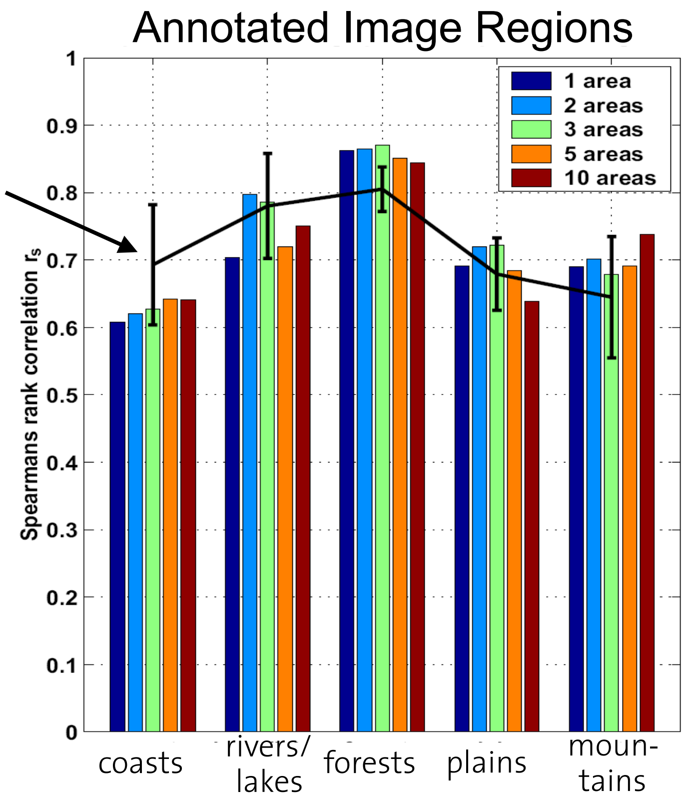
Prototype Approach + Perceptually Plausible Distance

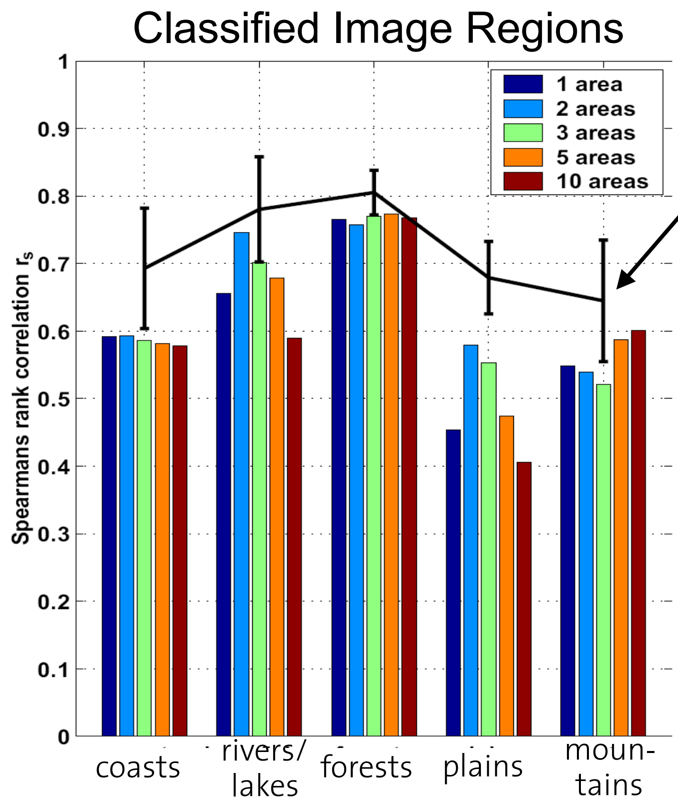
$$d^c_{PPD} = \sum_{j=1}^{N} w^c_j (COV_j - p^c_j)^2$$

where $\mathbf{p}^c$ = Prototype of category c ,
$\mathbf{w}^c$ = concept weights of category c.

Concept weights w^c_j learned from human data

Average Inter-rater correlation





Average Inter-rater correlation

Automatic Typicality Ranking

Qualitative Comparison: 50 images of all five categories

10 top-ranked images relative to mountains

Automatically obtained ranking: Classified image regions

Human ranking

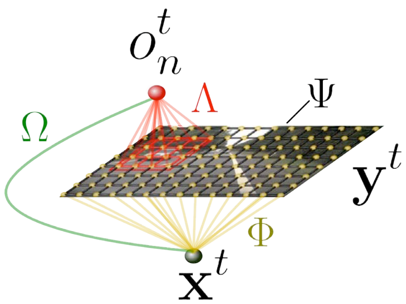
Quantitative comparison:

Spearman’s rank correlation between human and computational ranking.

Overview

- Semantic Scene Modeling [ijcv07,tap'06]
  - natural **scene categorization is not enough**
  - **aim for typicality ranking instead !**
  - joint work with **Julia Vogel**

- Joint Labeling of Objects and Scenes [eccv08]
  - dynamical conditional random field model
  - joint work with **Christian Wojek**



Joint Object and Scene Labeling: Motivation and Task Description

Input image

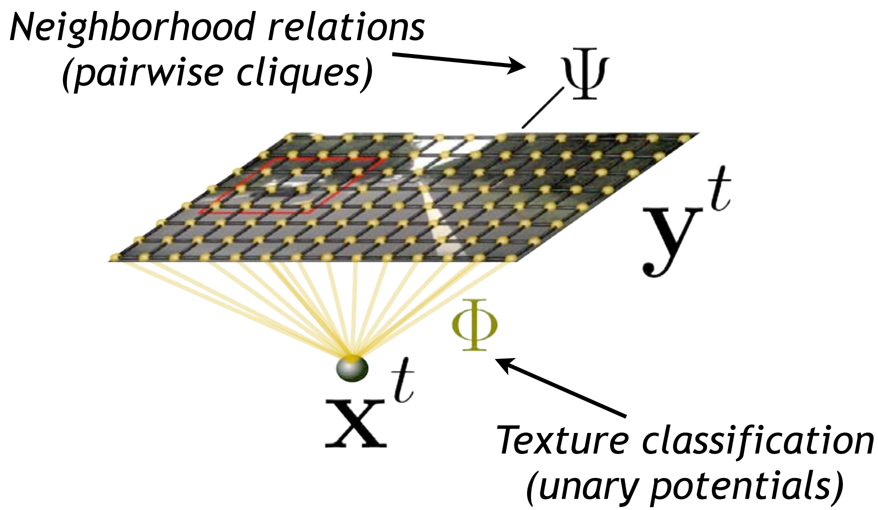
Desired Output
(Hand-labeled ground truth)

- Motivation:
  - Scene Labeling (=Context) supports object detection
  - Object detection supports scene labeling

- Approach:
  - 1. CRF for Scene Labeling
  - 2. Object-CRF to also include object detections
  - 3. Dynamic-Object-CRF to leverage temporal consistency

“Standard” Conditional Random Fields

- Conditional Random Field Models (CRFs) allow to model neighborhood relations
  - Unary Potentials
    - to label image regions locally (= nodes)
  - Edge potentials to model neighborhood relations
    - here: modeled with a logistic regression function
    - Parameters are learned via gradient descent in maximum likelihood setting
  - Loopy Belief Propagation used for inference



$$\log(P_{pCRF}(\mathbf{y}^t|\mathbf{x}^t, N_1, \Theta)) = \sum_i \Phi(y_i^t, \mathbf{x}^t; \Theta_\Phi) + \sum_{(i,j)\in N_1} \Psi(y_i^t, y_j^t, \mathbf{x}^t; \Theta_\Psi) - \log(Z^t)$$

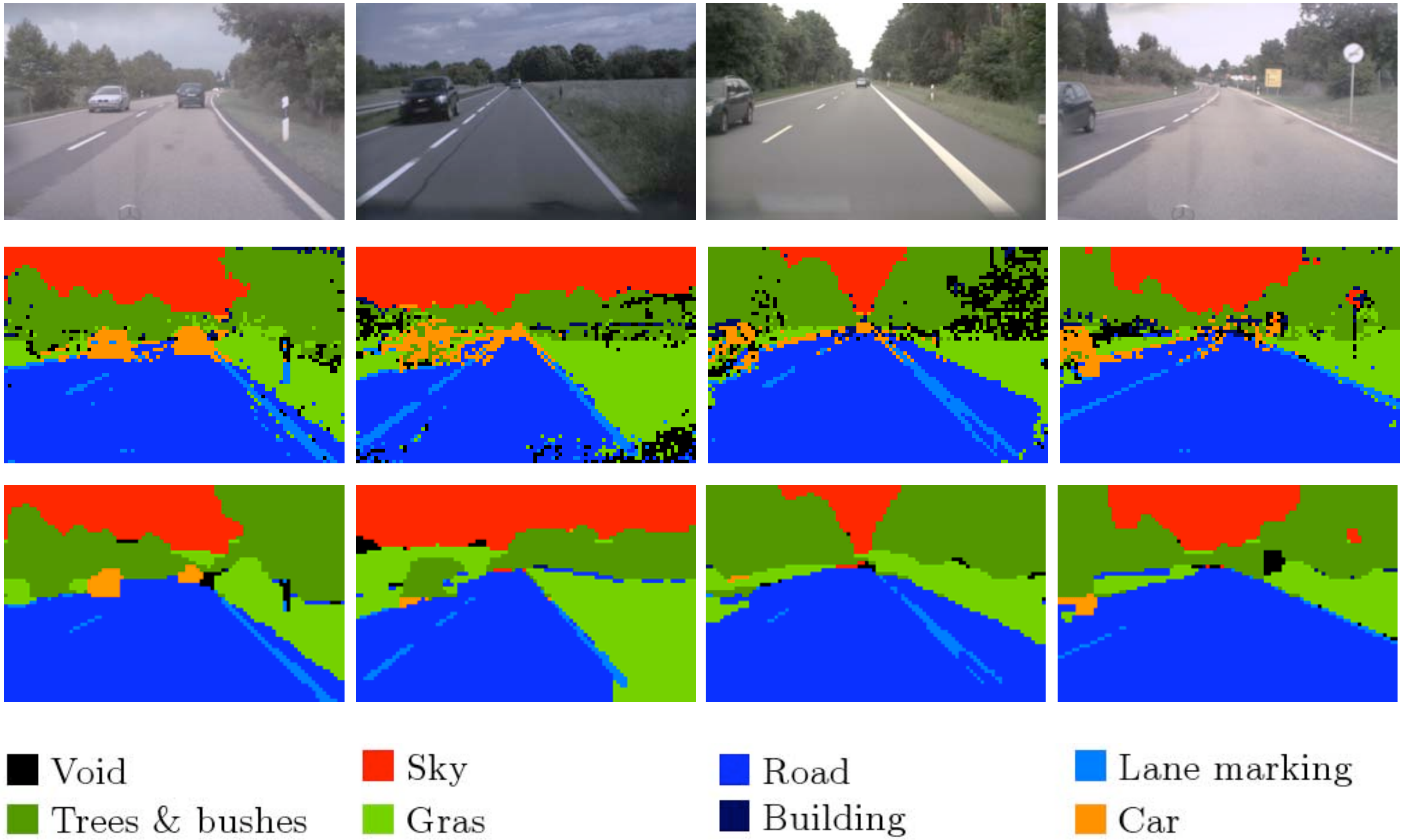
CRF for Scene Labeling

- Sample scene segmentations

Input

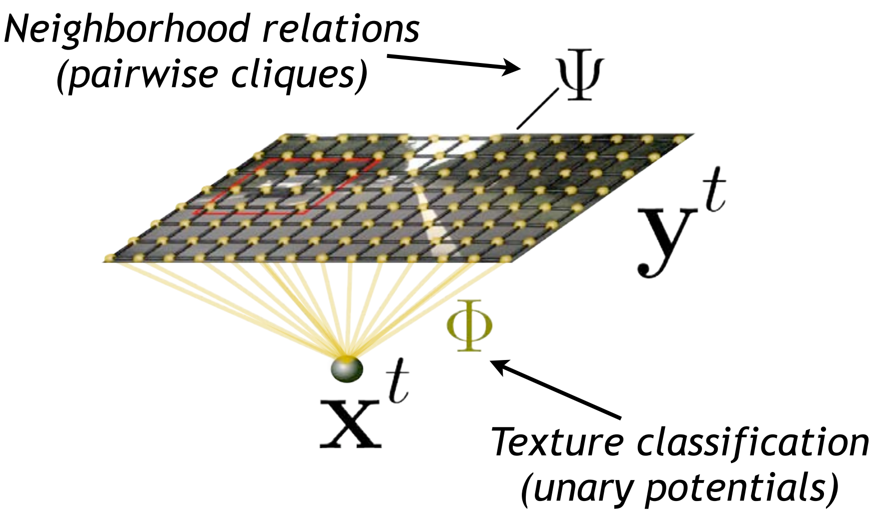
Unary Potentials

CRF with pairwise relations

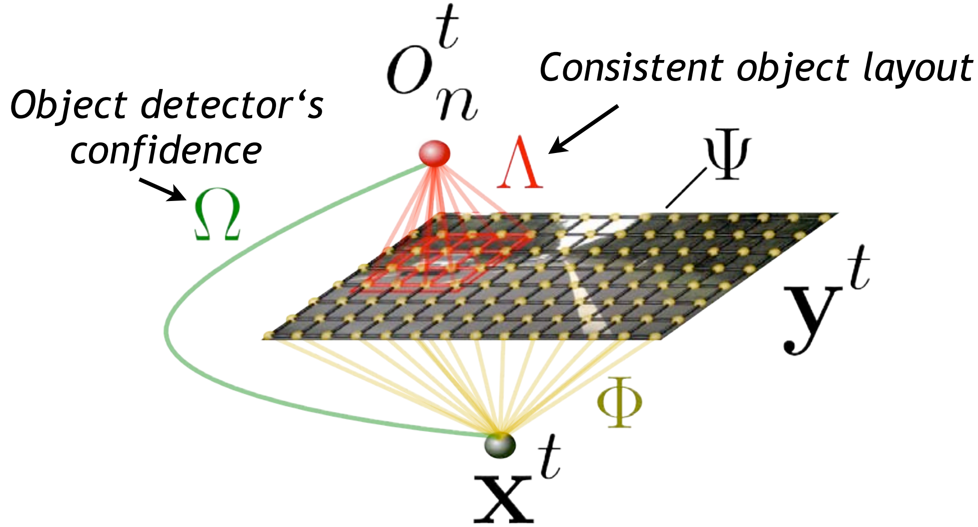


Object CRFs

Standard CRF for Scene Labeling

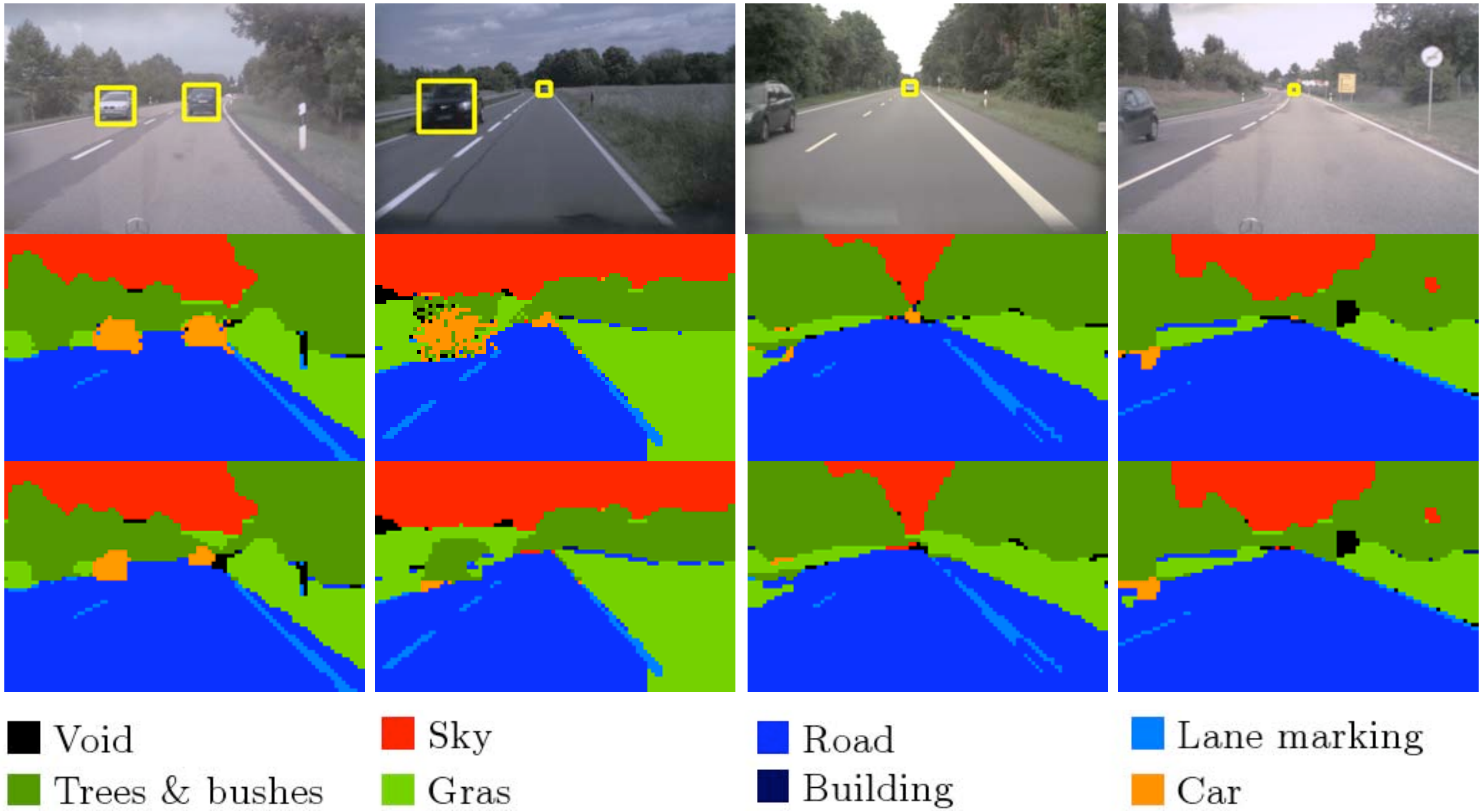


Object CRF



- Object CRF: Joint Labeling of Objects and Scene
  - Add additional nodes for each object hypothesis
    - Object detector's SVM margin is mapped to "pseudo probability" for the unary potential
    - Interaction weights model consistent object layout (Winn & Shotton CVPR'06)

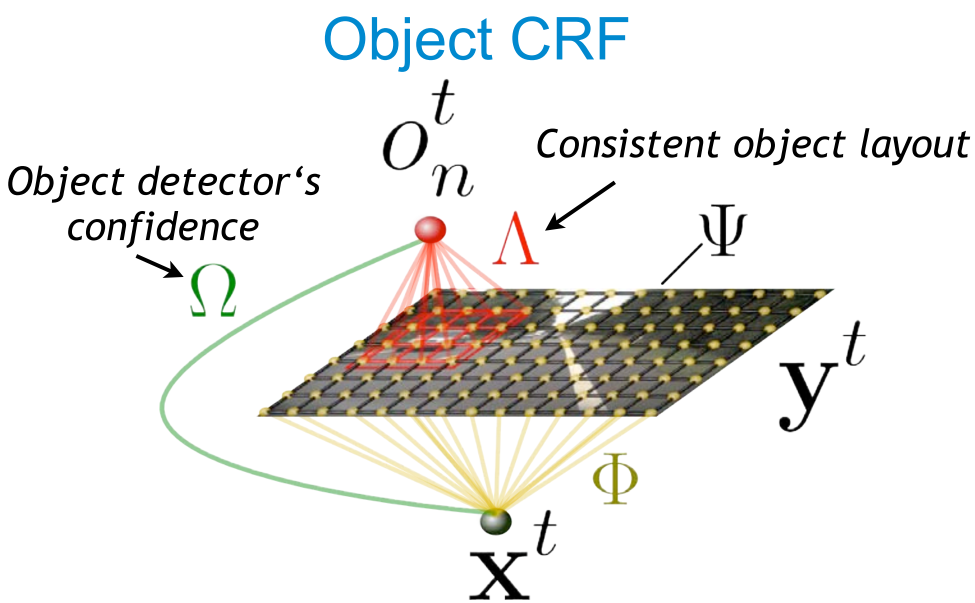
Object CRFs - Results

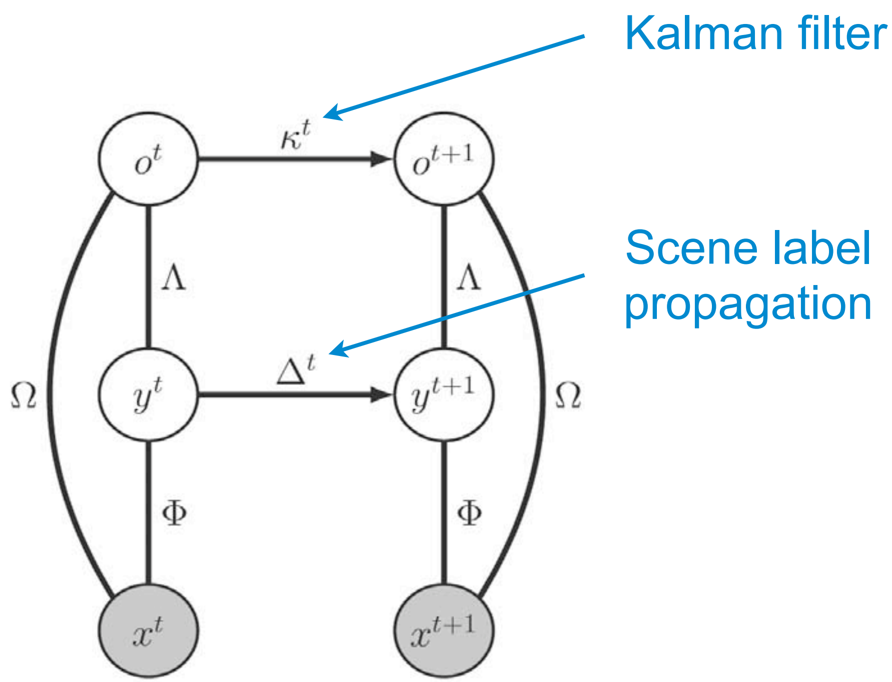


- Improvement for detected cars
- Small scale cars are segmented much better
- Segmentation on partially visible cars can still be improved

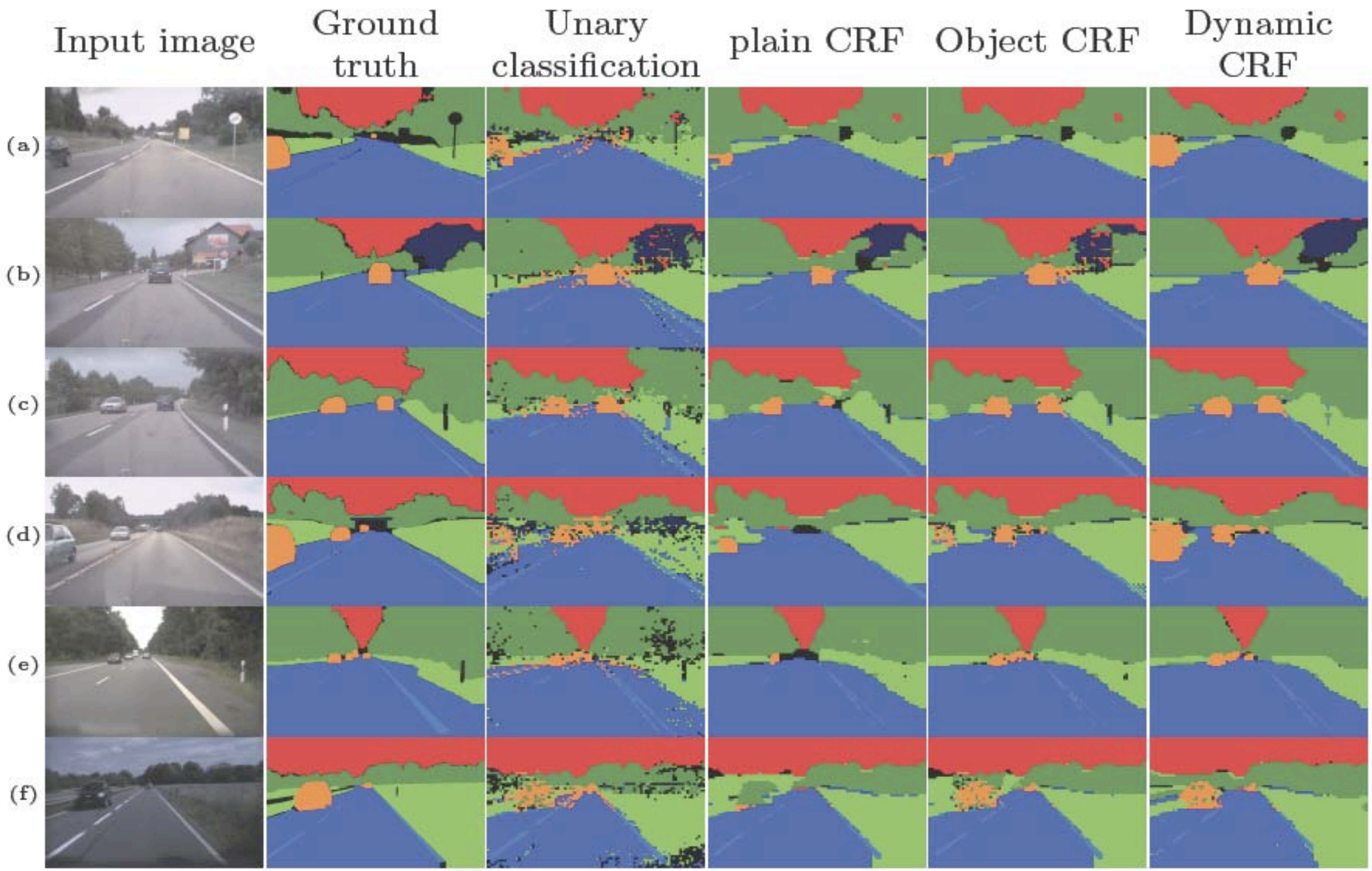
Dynamic CRFs

- Temporal integration
- Scene and Objects have different dynamics
  - object dynamics: track objects with a Kalman filter
  - scene dynamics: propagate scene labeling using odometry data





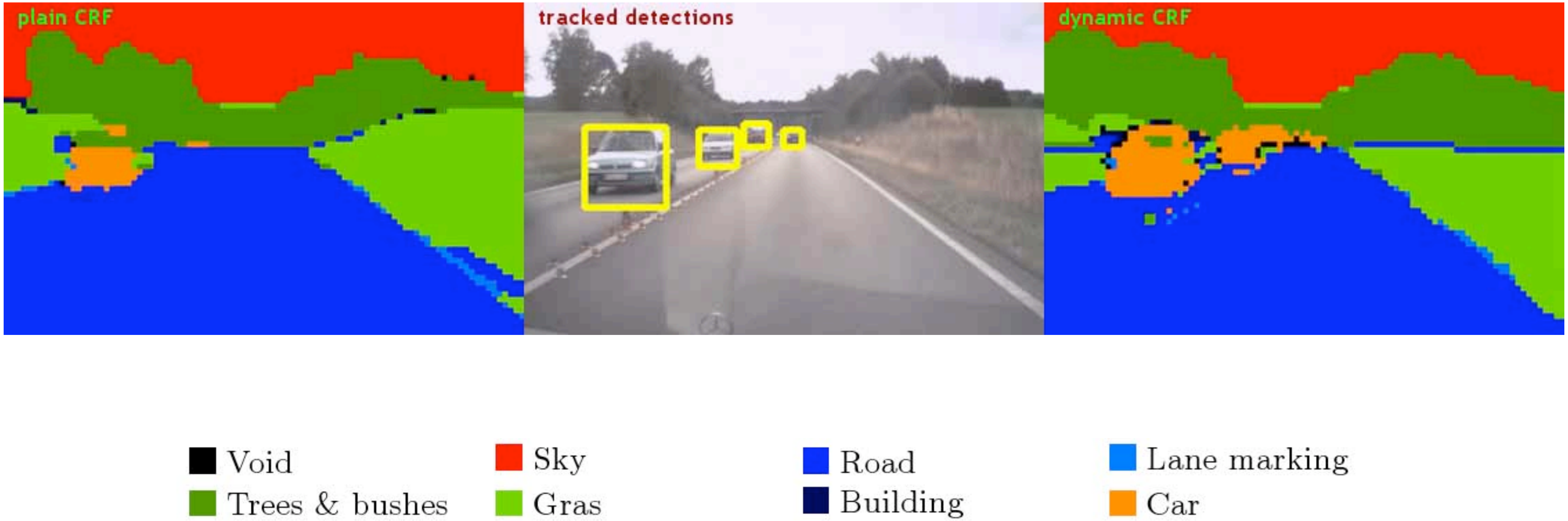
Results - Overview



Results – Video 1

Void	Sky	Road	Lane marking
Trees & bushes	Gras	Building	Car

Results – Video 2



Overview

- Semantic Scene Modeling [ijcv07,tap'06]
  - natural **scene categorization is not enough**
  - aim for typicality ranking instead !
  - joint work with **Julia Vogel**

- Joint Labeling of Objects and Scenes [eccv08]
  - dynamical conditional random field model
  - joint work with **Christian Wojek**