Object Detection in Crowded Scenes

Bastian Leibe
Computer Vision Laboratory
ETH Zurich

joint work with:
Nico Cornelis,
Kurt Cornelis,
Luc Van Gool,

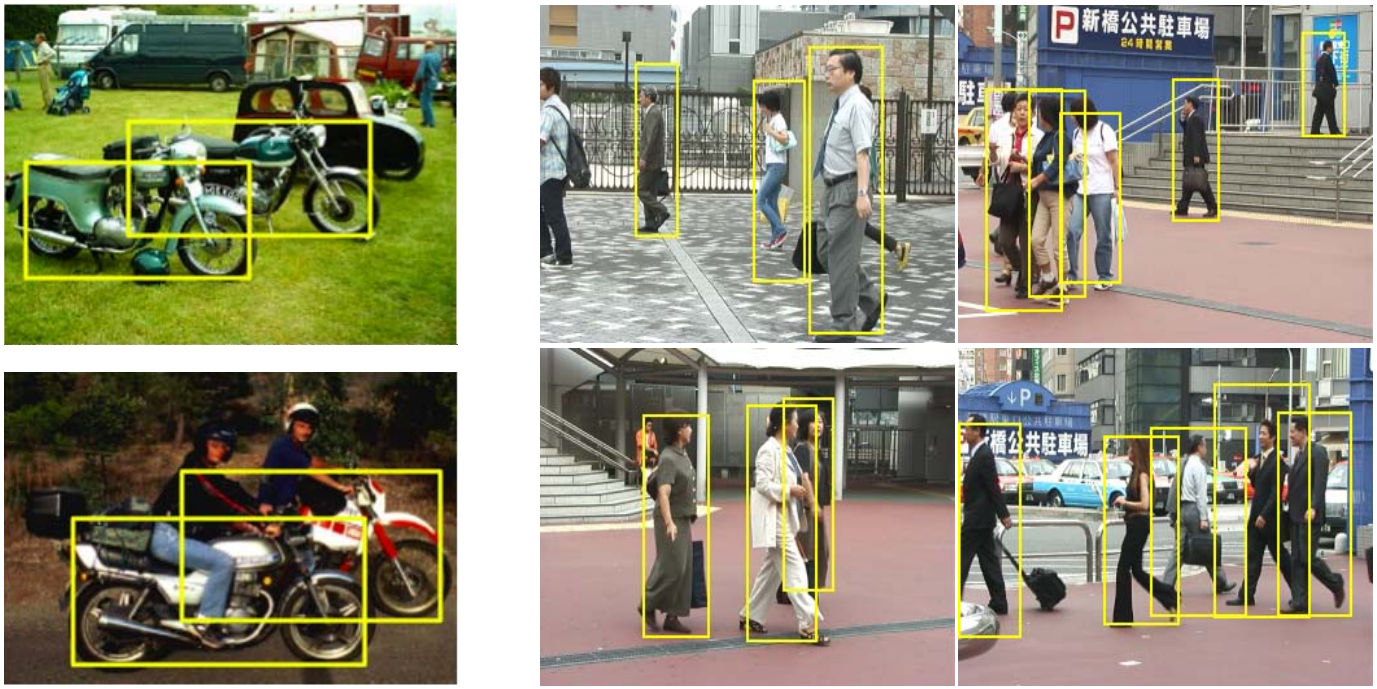
Edgar Seemann,
Krystian Mikolajczyk,
Bernt Schiele

CMP Seminar, Prague, 05.10.2006

CVPR'05, BMVC'06
CVPR'06 Video Proceedings
DAGM'06

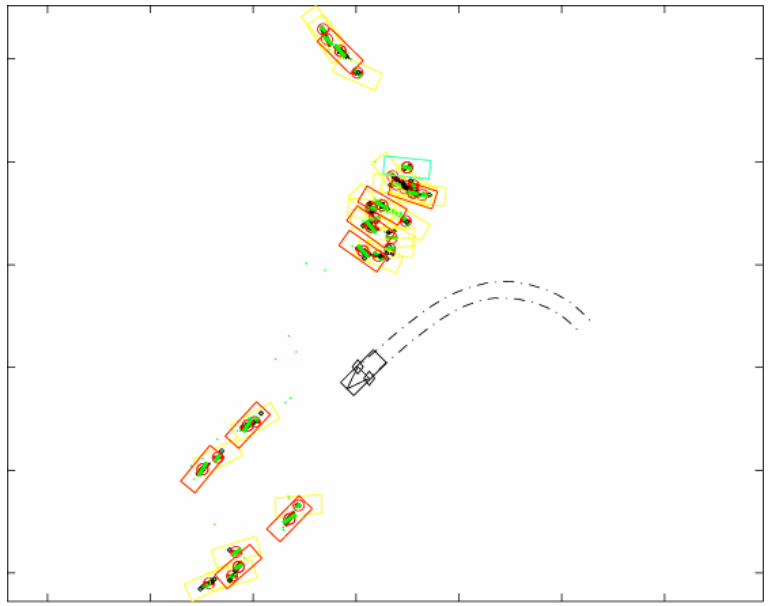
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Motivation



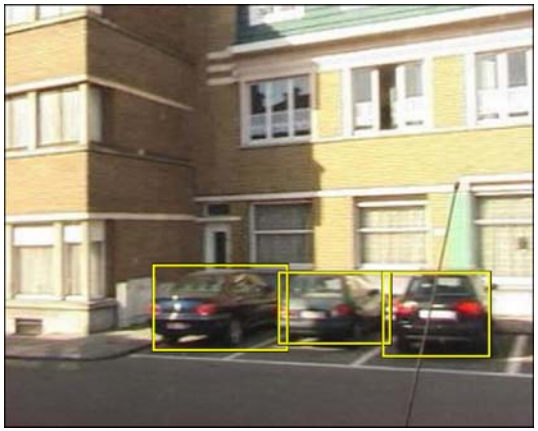
- **Object Detection in Crowded Scenes**
  - **Recognize and localize objects of a learned category**
  - **Learn object variability**
    - **Changes in appearance, scale, and articulation**
  - **Compensate for clutter, overlap, and occlusion**

Motivation (2)



- **Urban scene analysis from a moving vehicle**
  - **Detect objects in the image**
  - **Localize them in 3D**
  - **Build up a metric scene model**
- **Applications e.g. in driver assistance systems**

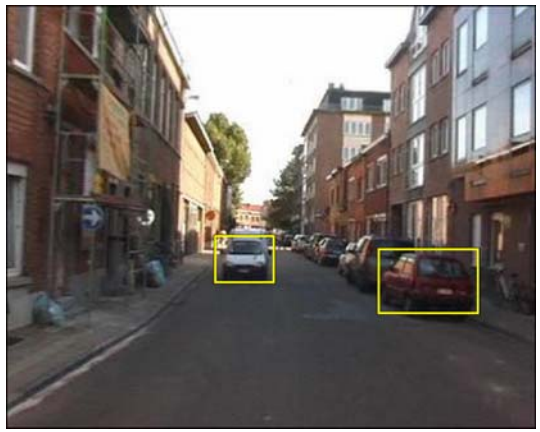
Challenges



Motion blur

Brightly-lit areas

Lense flaring



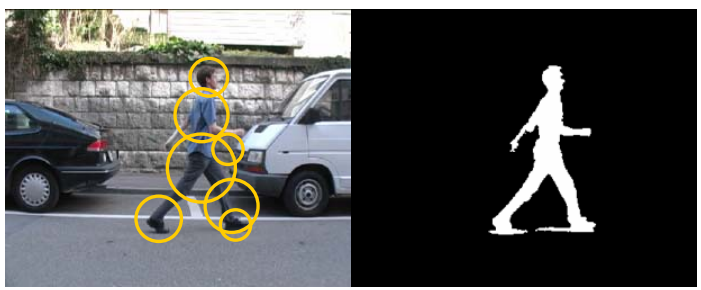
Dark shadows

+ Intra-category variability, multiple viewpoints, partial occlusion, …

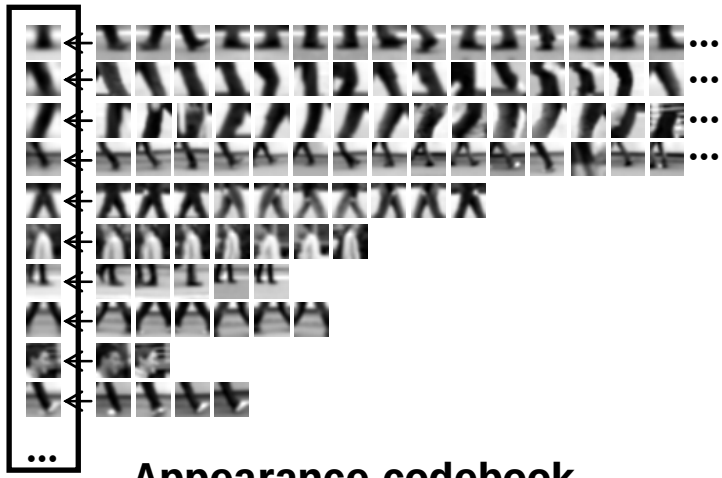
Outline

- **Object detection approach**
  - Cognitive Loop between recognition and segmentation
  - Hypothesis Generation → Segmentation → Verification
- **Recent extensions**
  - Evaluation of different local features
  - Multi-cue integration
- **Application for urban scene analysis**
  - Cognitive Loop between recognition and 3D reconstruction
  - Real-time scene geometry estimation
  - Object detection, 3D localization, and temporal integration
  - Feedback into geometry estimation

Implicit Shape Model - Representation

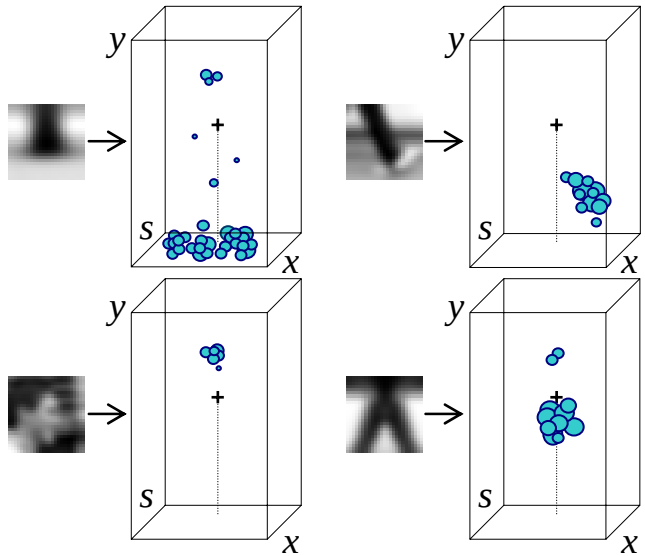


training images
(+reference segmentation)



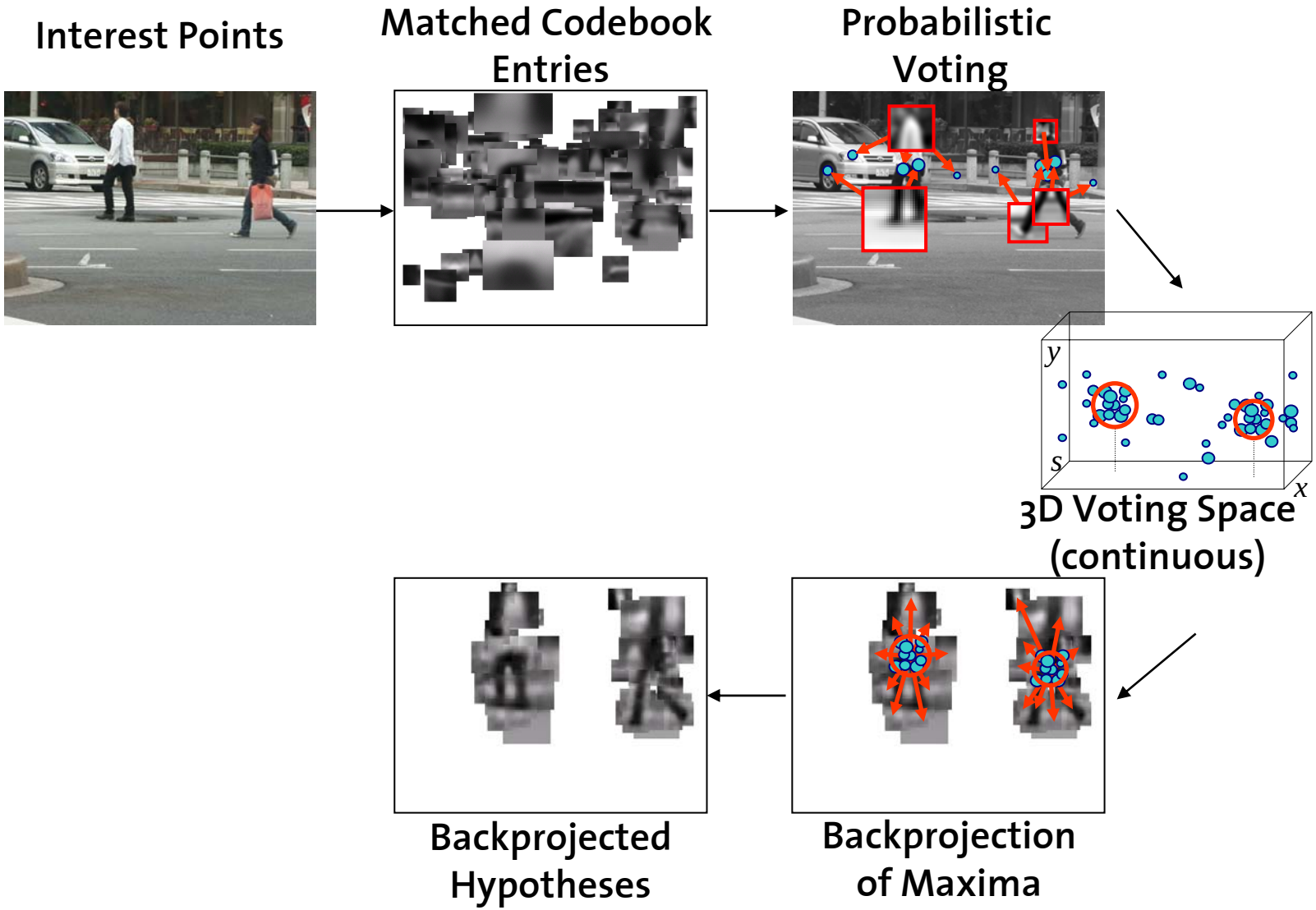
Appearance codebook

- **Learn appearance codebook**
  - Extract patches at interest points
  - Agglomerative clustering ⇒ codebook
- **Learn spatial distributions**
  - Match codebook to training images
  - Record matching positions on object



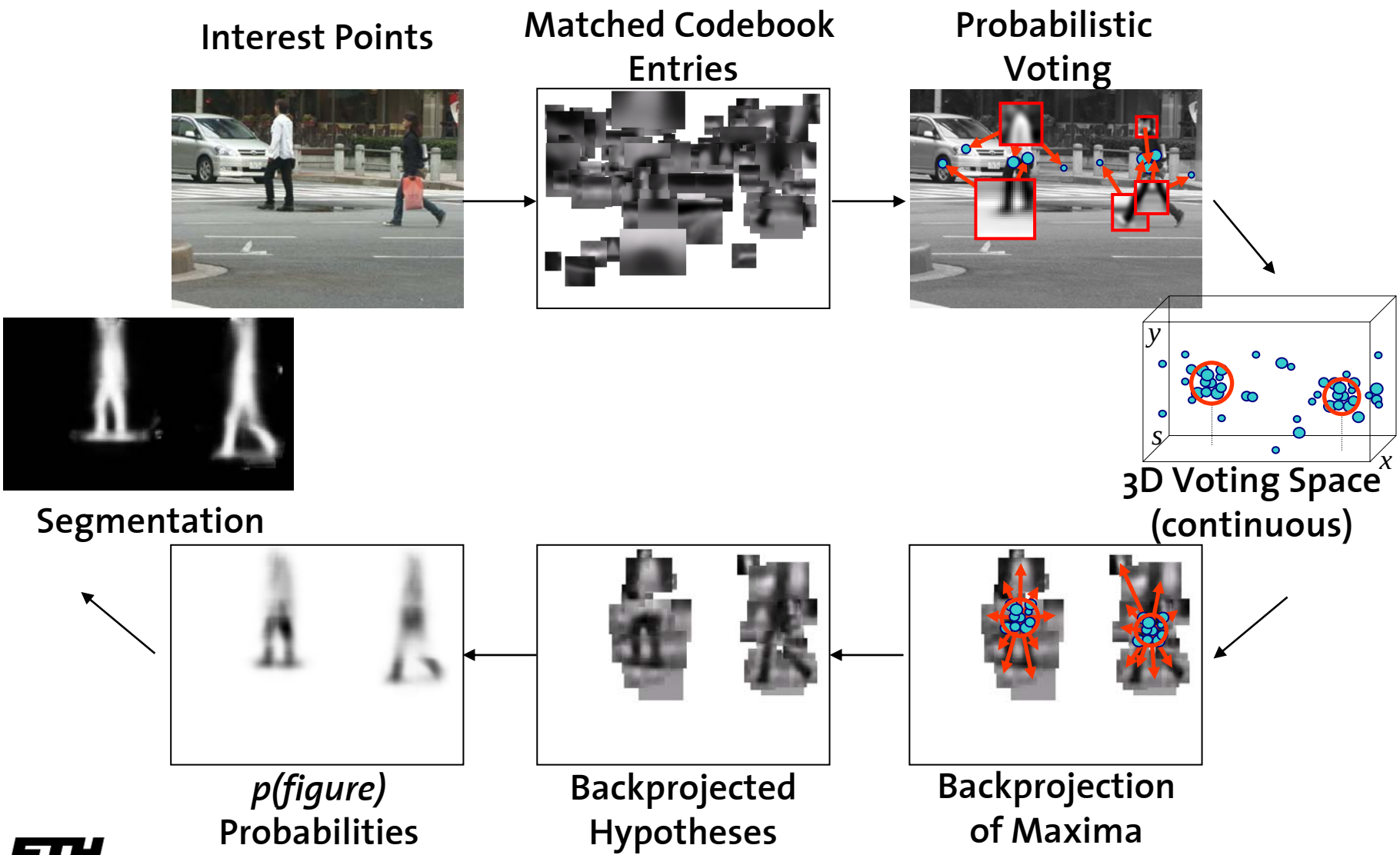
Spatial occurrence distributions

Implicit Shape Model - Recognition

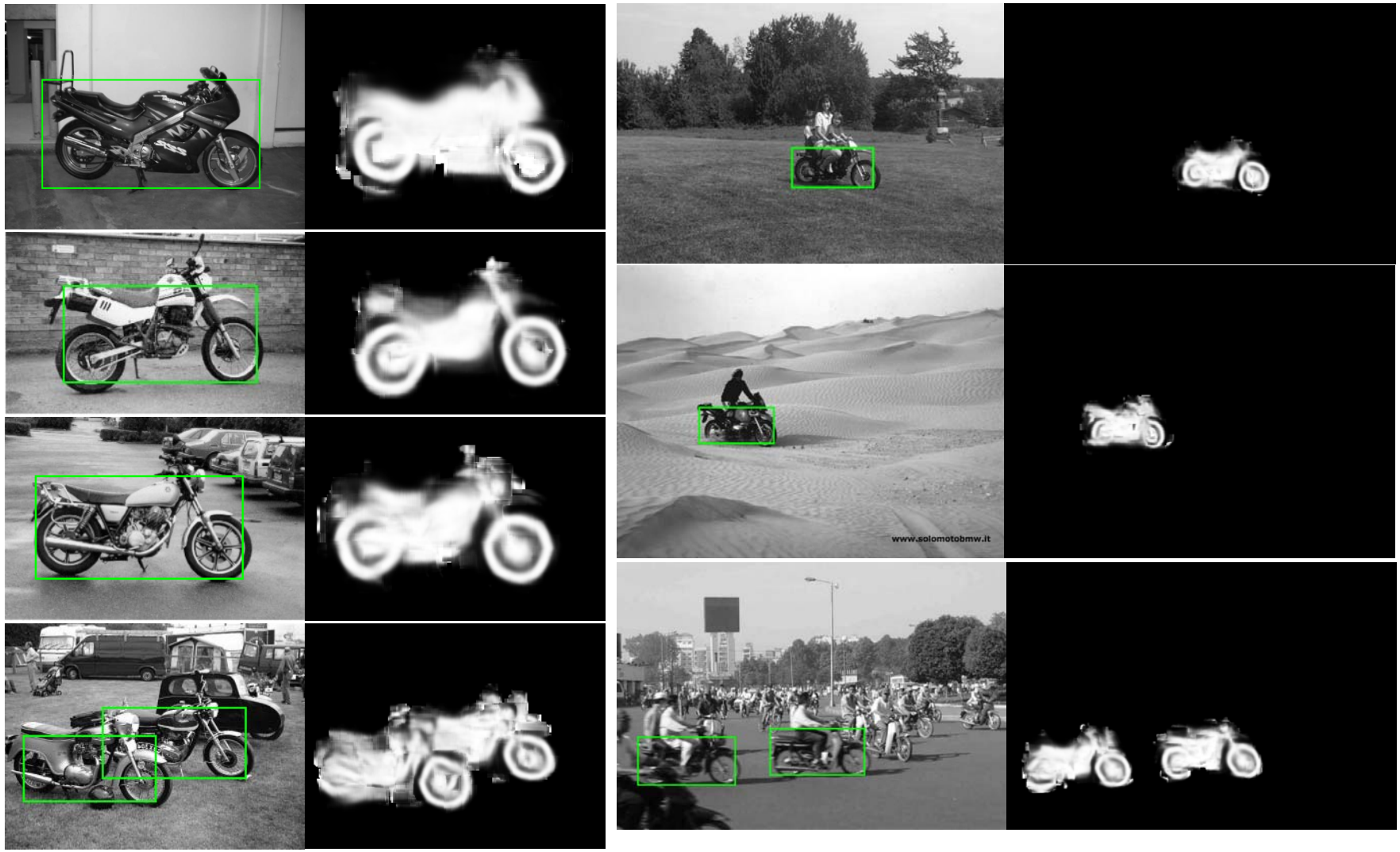


[Leibe & Schiele,04]

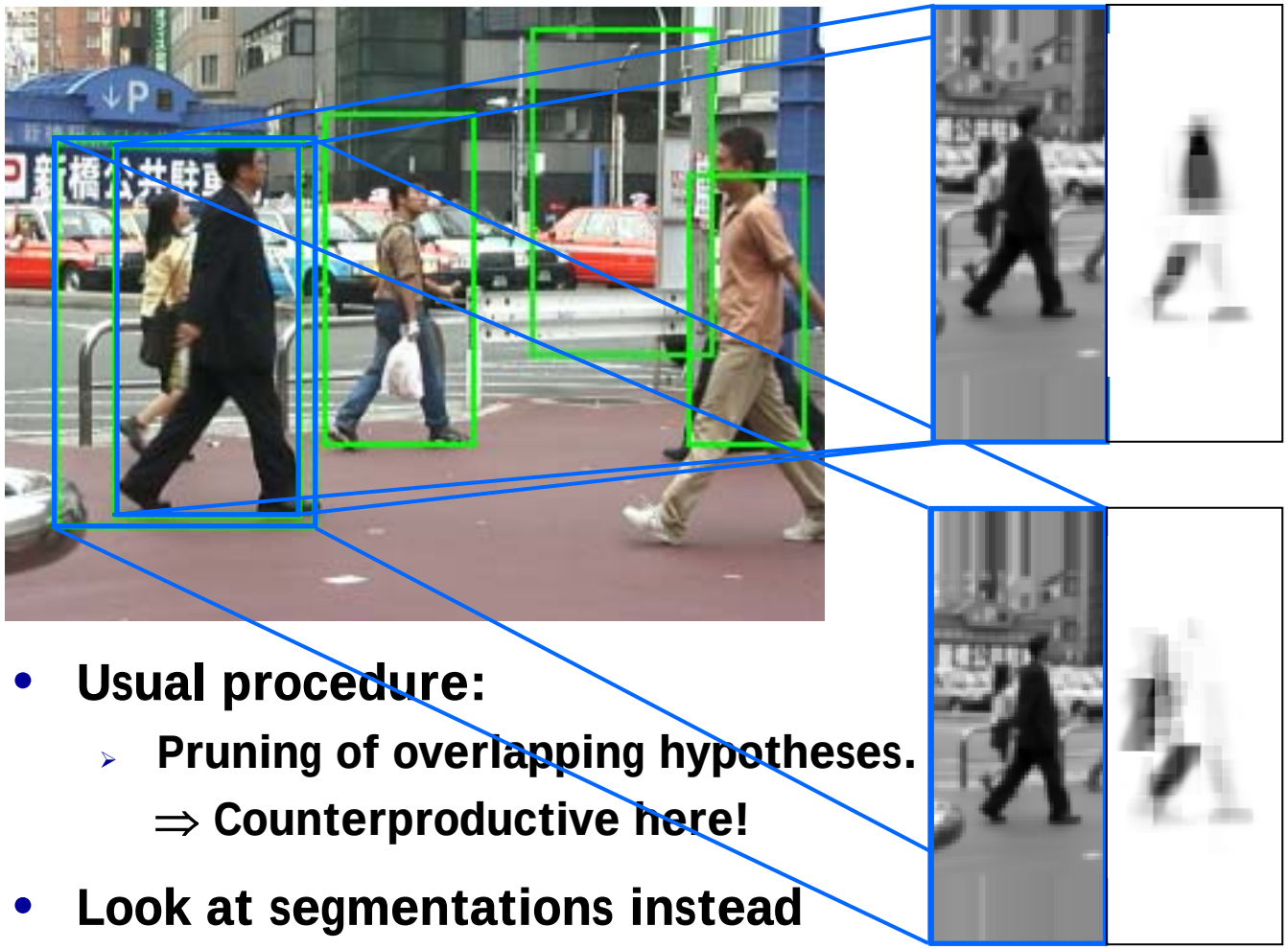
Implicit Shape Model - Recognition



Example Results: Motorbikes



Segmentation-Based Verification



- **Usual procedure:**
  - **Pruning of overlapping hypotheses.**
    ⇒ Counterproductive here!
- **Look at segmentations instead**
  ⇒ Support comes from different areas!

Formalization in MDL Framework

- **Savings of a hypothesis**

$$S_h = K_0 S_{area} - K_1 S_{model} - K_2 S_{error}$$

- **with**
  - S_{area} **: #pixels** N **in segmentation**
  - S_{model}**: model cost, proportional to** *expected area* A_s
  - S_{error} **: estimate of error, according to**

$$S_{error} = \sum_{\mathbf{p} \in Seg(h)} (1 - p(\mathbf{p} = figure|h))$$

- **Final form of equation**

$$S_h = -\frac{K_1}{K_0} + (1 - \frac{K_2}{K_0})\frac{N}{A_s} + \frac{K_2}{K_0}\frac{1}{A_s} \sum_{\mathbf{p} \in Seg(h)} p(\mathbf{p} = fig.|h)$$

Formalization in MDL Framework (2)

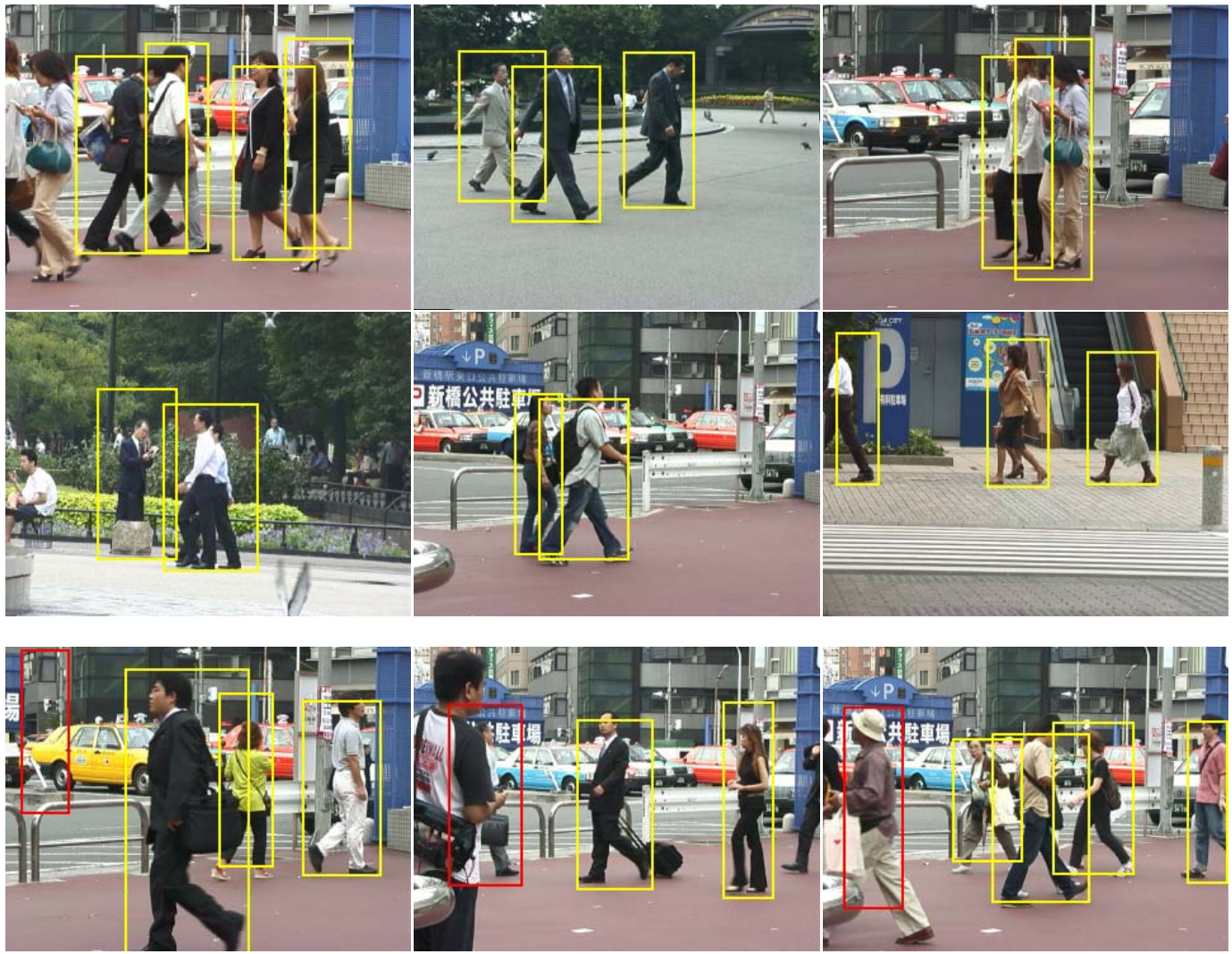
- **Savings of *combined* hypothesis**

$$S_{h_1 \cup h_2} = S_{h_1} + S_{h_2} - S_{area}(h_1 \cap h_2) + S_{error}(h_1 \cap h_2)$$

- **Goal: Find combination that best explains the image**
  - **Quadratic Boolean Optimization problem** [Leonardis et al,95]

$$S(\hat{m}) = \max_m m^T Q m = \max_m m^T \begin{bmatrix} S_{h_1} & \cdots & \frac{1}{2} S_{h_1 \cap h_N} \\ \vdots & \ddots & \vdots \\ \frac{1}{2} S_{h_1 \cap h_2} & \cdots & S_{h_N} \end{bmatrix} m$$

Qualitative Results

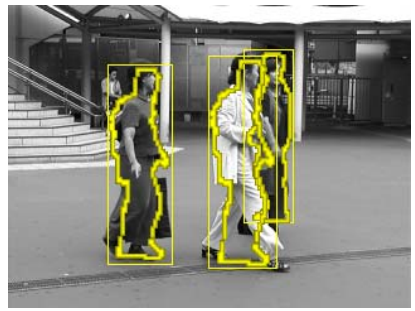


[Leibe, Seemann, Schiele, CVPR'05]

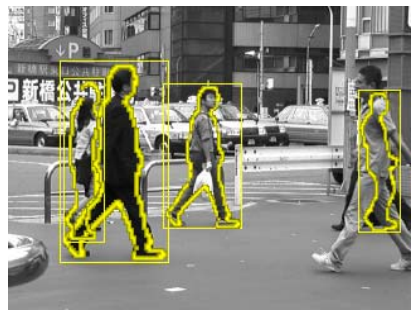
Estimated Articulations

Problem cases

Walking direction



Body pose



Detection scale

Detections at EER

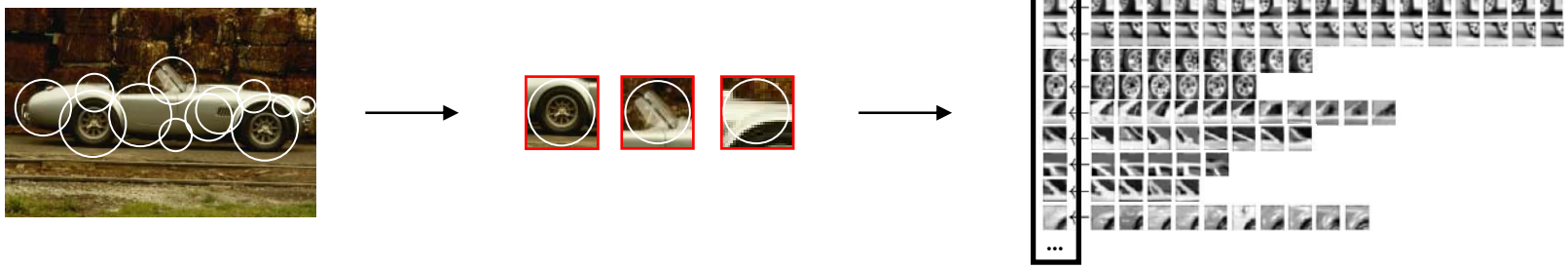
Single-frame recognition – no temporal continuity used!

[Leibe, Seemann, Schiele, CVPR'05]

Outline

- **Recent extensions**
  - **Evaluation of different local features**
  - **Multi-cue integration**

Which Features to Use?



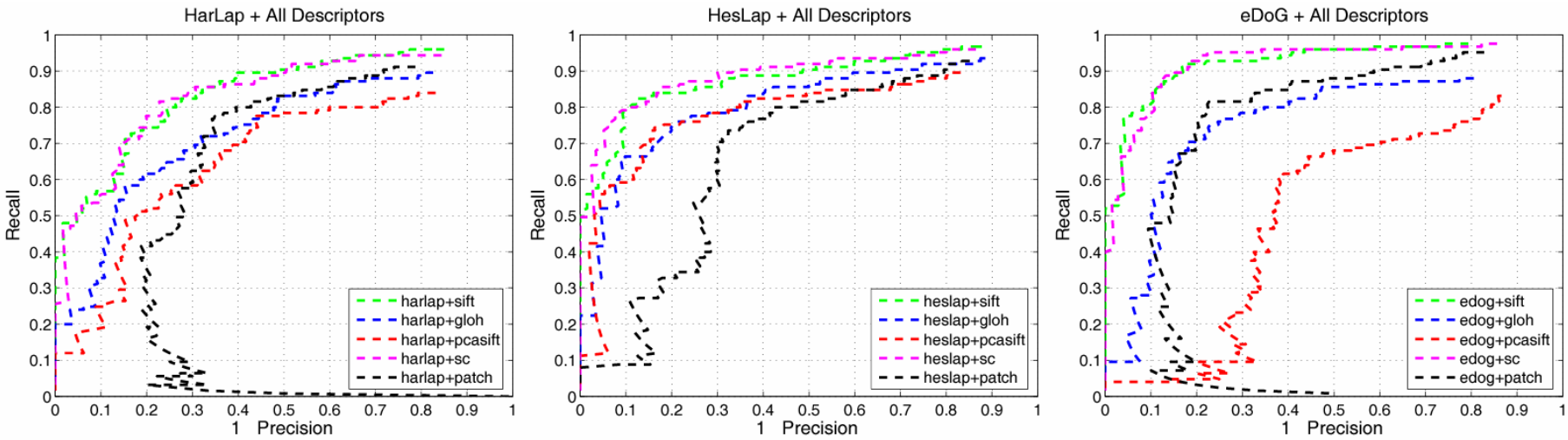
Region Detector
- DoG
- Harris Laplace
- Hessian Laplace
- (Salient Regions)
- (MSER)

Region Descriptor
- Patches
- SIFT
- GLOH
- PCA-SIFT
- Shape Context

Cue Codebook
(for each combination)

[Leibe, Mikolajczyk, Schiele, BMVC'06]

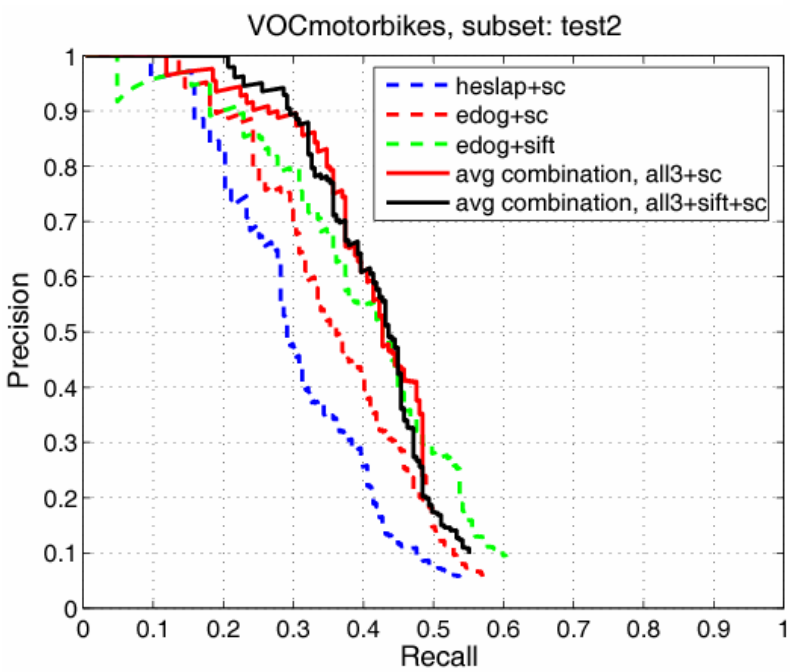
Local Feature Evaluation

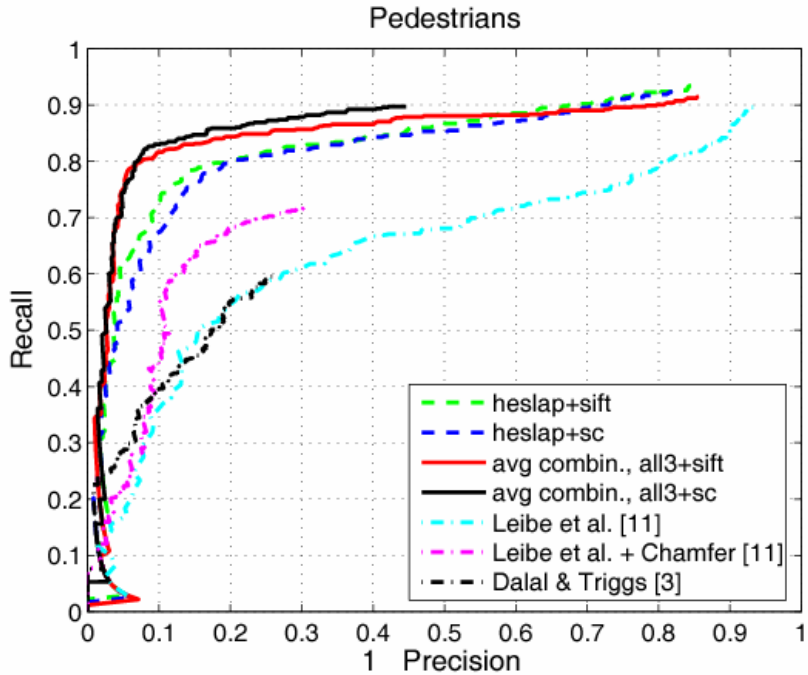


- **Results on TUD Motorbike set (115 imgs, 125 objects)**
  - **Best descriptors: SIFT + Shape Context**
  - **Best detectors: Hes-Lap + DoG**
  - **Performance significantly improved compared to DoG+Patches**

[Leibe, Mikolajczyk, Schiele, BMVC'06]

Results on Other Data Sets





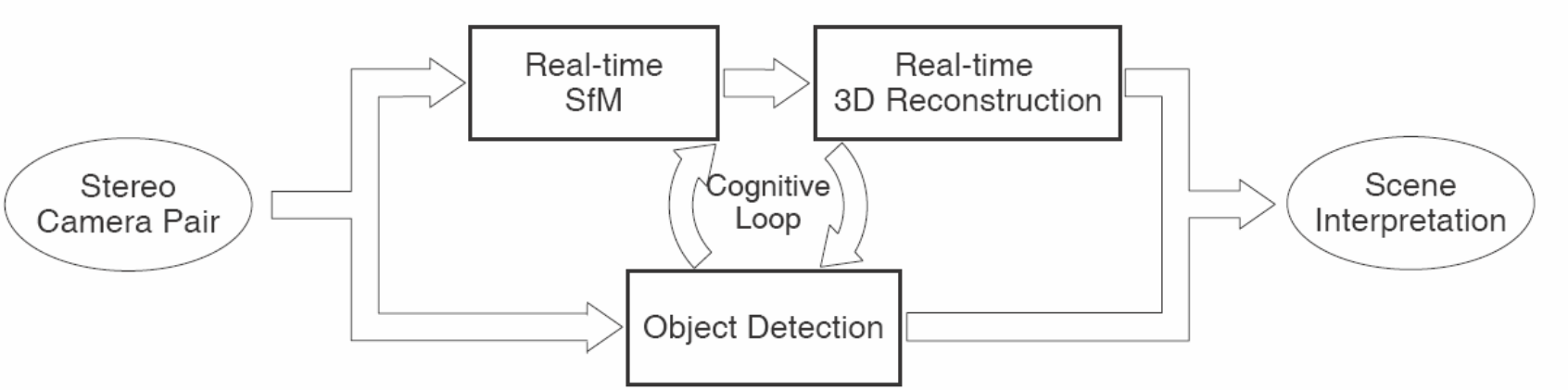
- **Further improvements by combining several local cues**
  - **Improved recognition performance**
  - **Increased recall at low false-positive rates**

  ⇒ **Reaching a level where the system can be used for applications.**

Outline

- **Application for urban scene analysis**
  - **Cognitive Loop between recognition and 3D reconstruction**
  - **Real-time scene geometry estimation**
  - **Object detection, 3D localization, and temporal integration**
  - **Feedback into geometry estimation**

Cognitive Loop with 3D Geometry

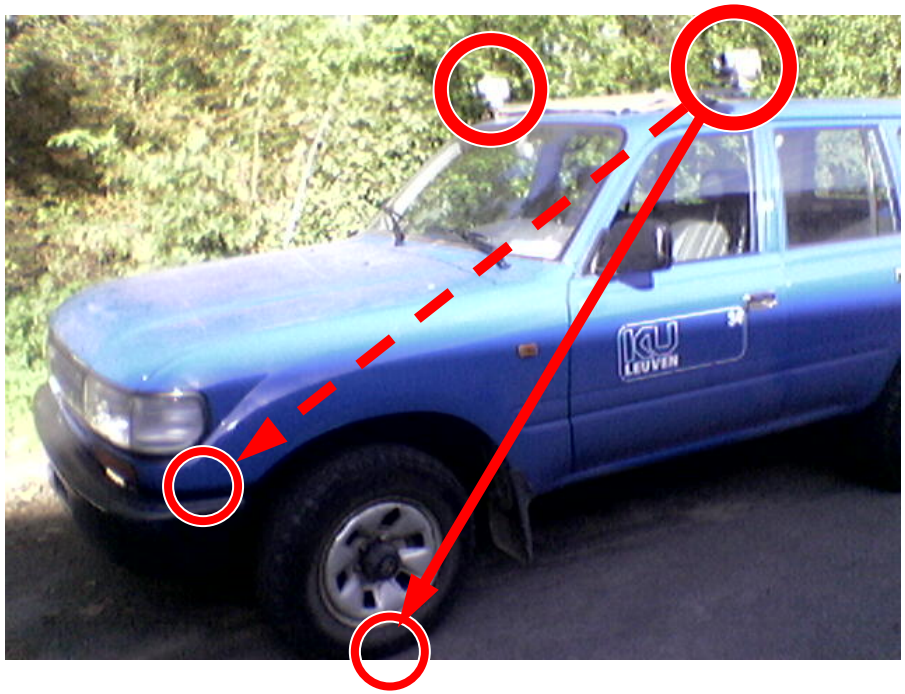


- **Connect recognition and reconstruction**
- **Reconstruction pathway delivers scene geometry**
  - ⇒ **Greatly improves recognition performance**
- **Recognition detects objects that disturb reconstruction**
  - ⇒ **More accurate geometry estimate**

Outline

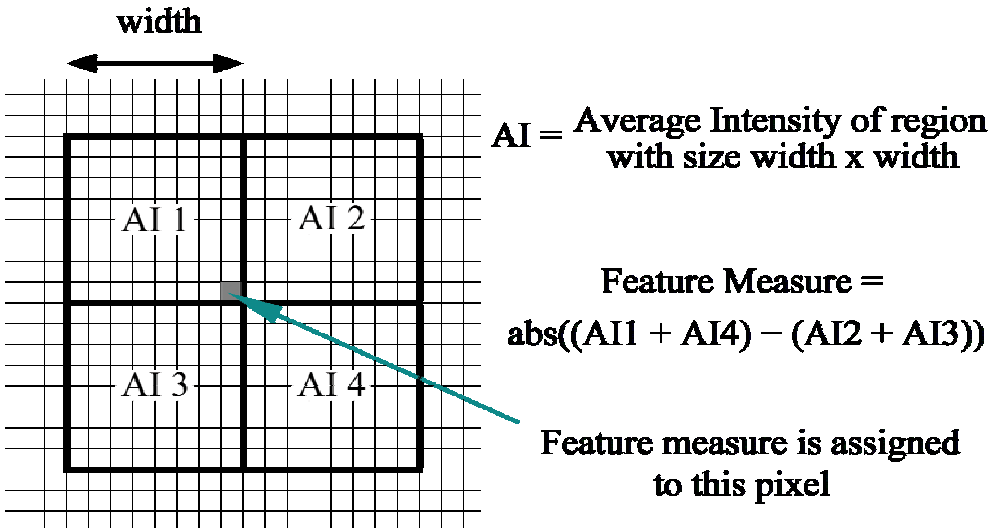
- Hardware setup
- Reconstruction pathway
  - Real-time Structure-from-Motion
  - Real-time dense reconstruction
- Recognition pathway
  - Local-feature based object detection
  - Incorporation of scene geometry
  - Temporal integration in world coordinate frame
  - Feedback to reconstruction
- Results and Conclusion

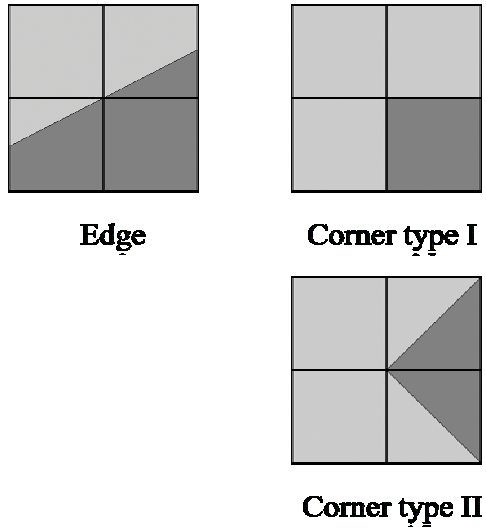
Hardware Setup



- Stereo camera rig mounted on top of the vehicle
- Calibrated w.r.t. wheel base points
- Video streams captured at 25 fps, 360×288 resolution

Real-Time Structure-from-Motion

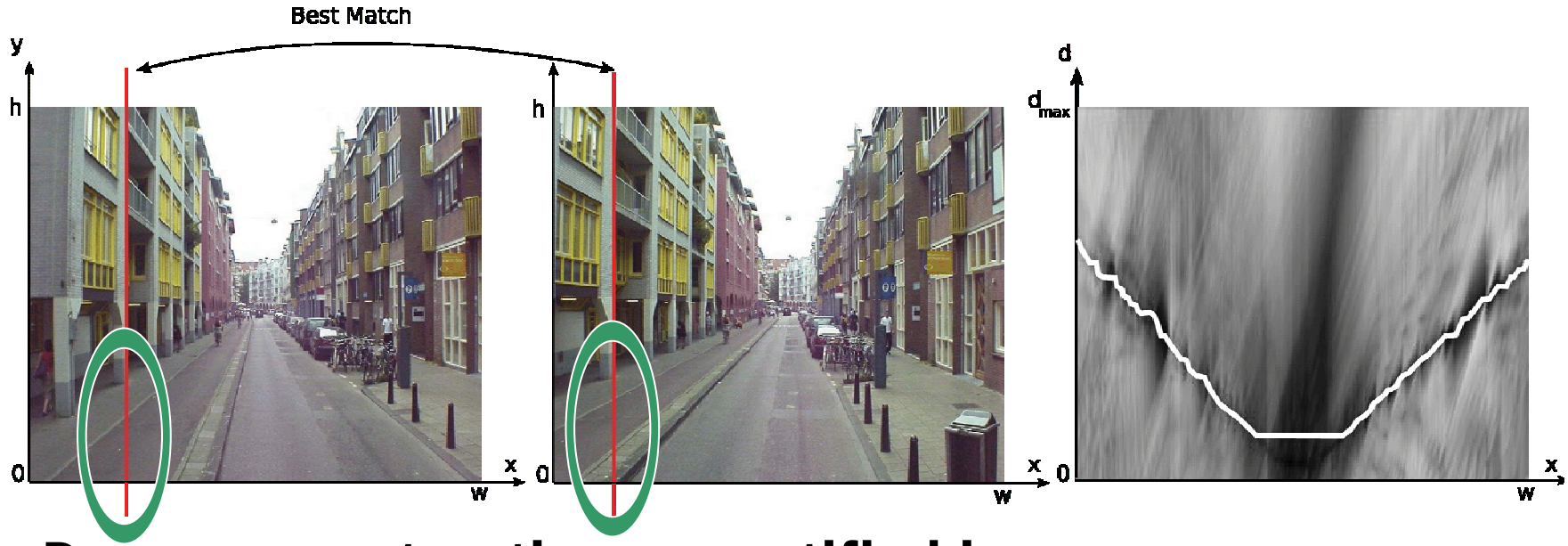




- **Basis: very fast feature matching**
  - **Simple features**
  - **Optimized for urban environment**
  - **Only computed on green channel of a single camera**
- **Rest: standard SfM pipeline**

[Cornelis et al., CVPR'06]

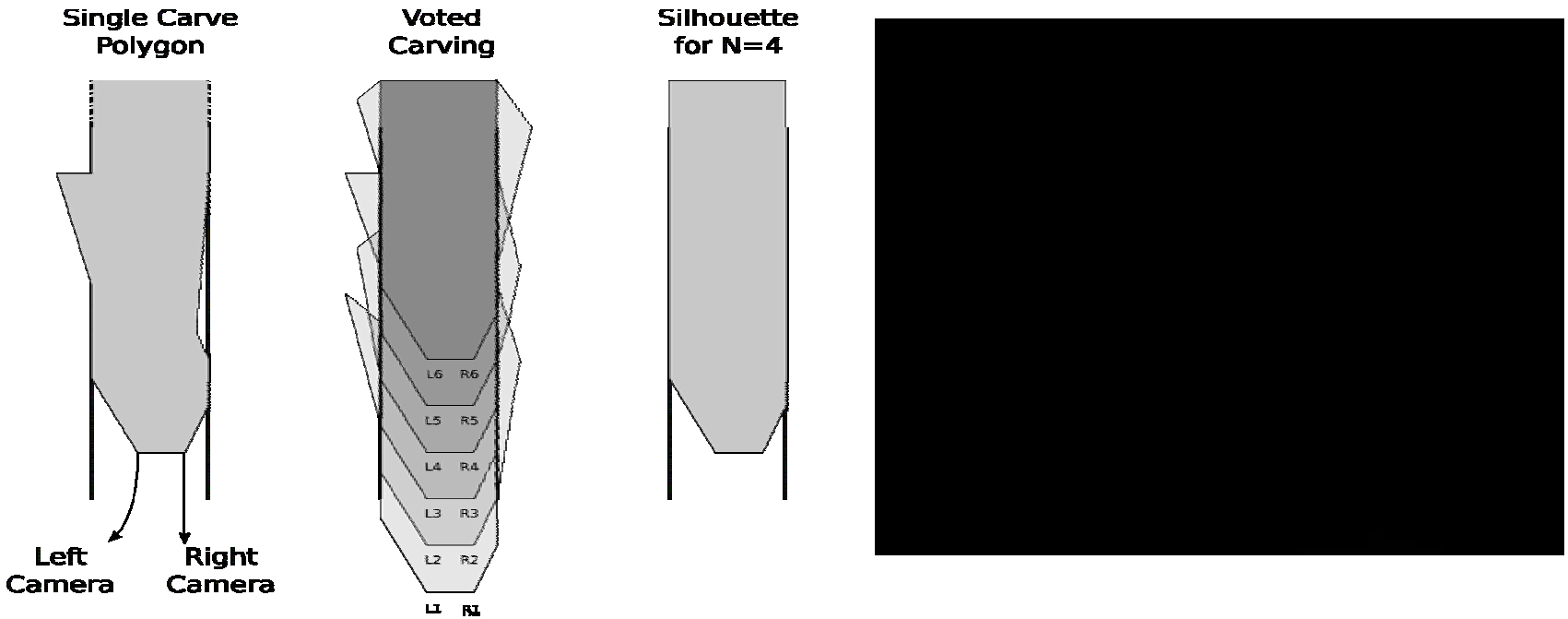
Real-Time Dense Reconstruction



- **Dense reconstruction on rectified images**
  - *Ruled surface* **assumption to speed-up dense reconstruction**
  - **Correlation measure: Sum of per-pixel SSDs along vertical lines**
  - **Line-sweep algorithm with ordering constraints (DP)**
  - **Fast computation on GPU**
- **Errors introduced by pixels not belonging to facades!**

[Cornelis et al., CVPR'06]

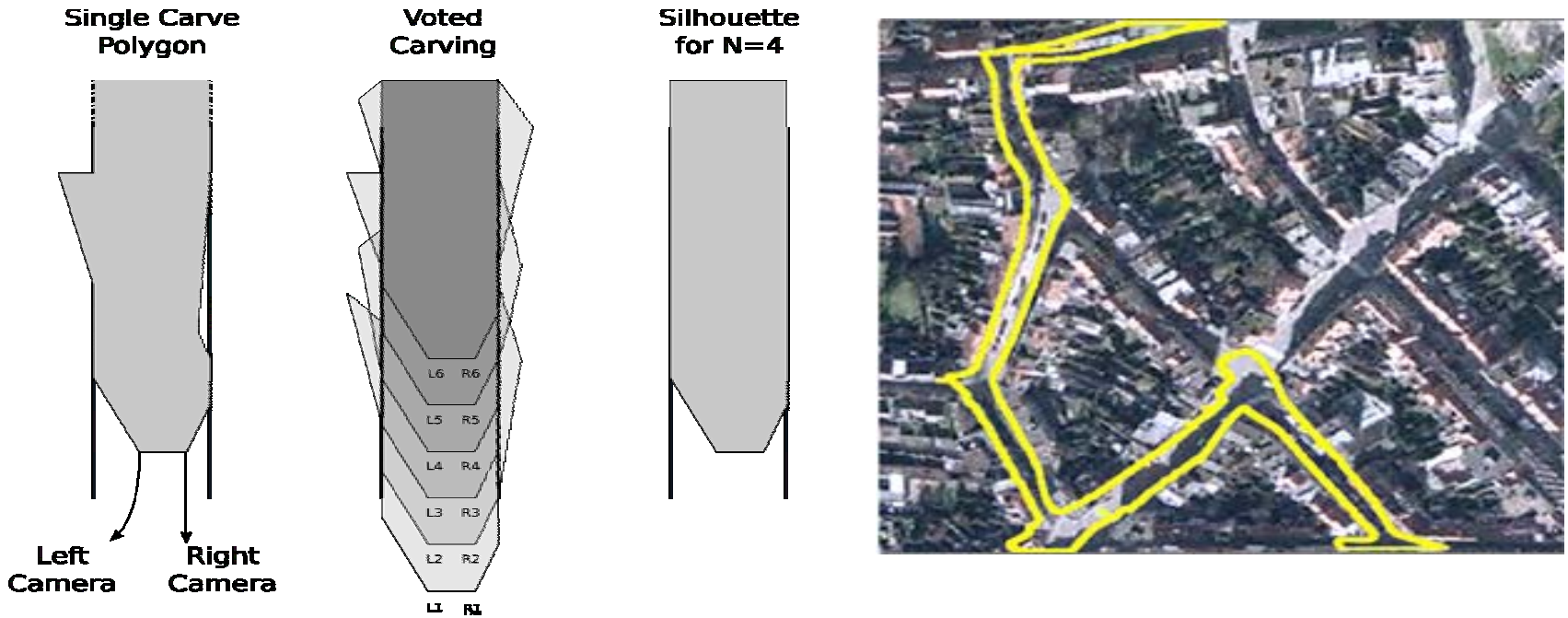
Real-Time Dense Reconstruction (2)



- **Merge dense reconstructions using known camera poses.**
- **"Voted polygon carving" on 2D projection**

[Cornelis et al., CVPR'06]

Real-Time Dense Reconstruction (2)



- **Merge dense reconstructions using known camera poses.**
- **"Voted polygon carving" on 2D projection**
- **Surfaces registered on world map using GPS**

[Cornelis et al., CVPR'06]

Textured 3D Model

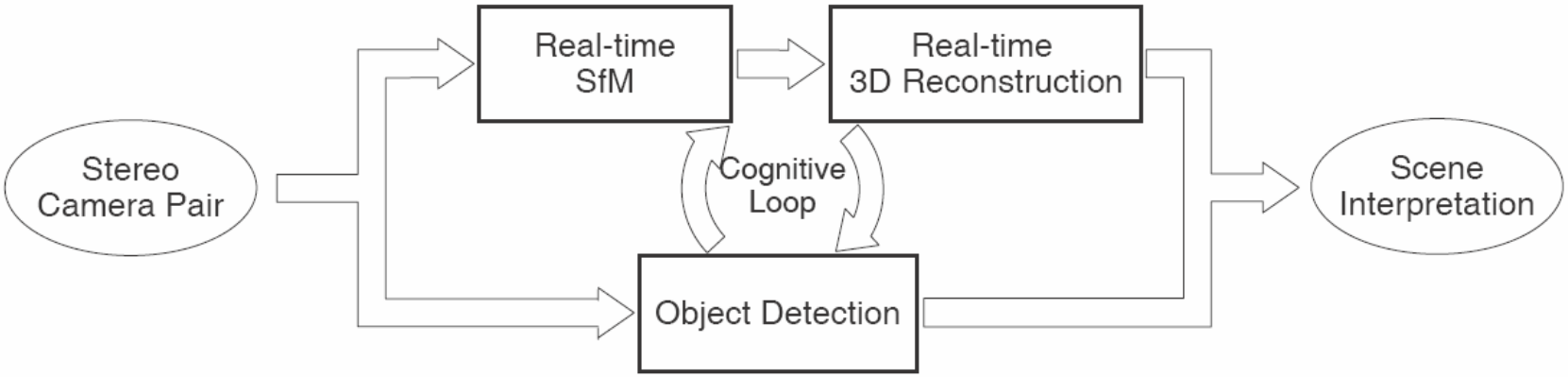
Original

3D Reconstruction

- **Run-times**
  - SfM + Bundle adjustment: 26-30 fps on CPU
  - Dense reconstruction: 26 fps on GPU

[Cornelis et al., CVPR'06]

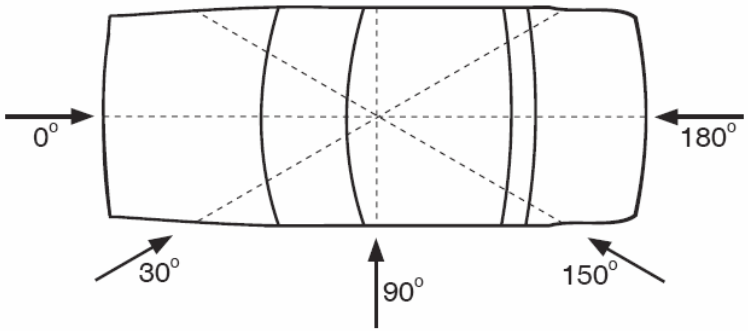
Information Flow into Recognition



- **For each frame, 3D reconstruction delivers**
  - **External camera calibration**
  - **Ground plane estimate**

  ⇒ **Used for improving recognition** *of the next frame.*

Appearance-Based Car Detection



	#images	mirrored
0°	117	
30°	138	X
90°	50	
150°	119	X
180°	119	

- **Bank of 5 single-view ISM detectors**
- **Each based on 3 local cues**
  - **Harris-Laplace, Hessian-Laplace, and DoG interest regions**
  - **Local Shape Context descriptors**
- **Semi-profile detectors additionally mirrored**
- **Not real-time yet…**

[Leibe, Mikolajczyk, Schiele, BMVC'06]

2D/3D Interactions

- **Likelihood of 3D hypothesis H given image I and 2D detections h:**

$$p(H|I) = \sum_h p(H|h,I)p(h|I) \sim \sum_h p(h|H)p(H)\boxed{p(h|I)}$$

recognition score (2D)

- **2D recognition score**
  - **Expressed in terms of per-pixel $p(figure)$ probabilities**

$$p(h|I) = \sum_{\mathbf{p}\in I} p(h|\mathbf{p}) = \sum_{\mathbf{p}\in Seg(h)} p(\mathbf{p} = figure|h)p(h)$$

2D/3D Interactions

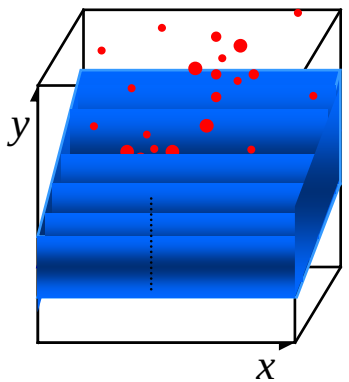
- **Likelihood of 3D hypothesis H given image I and 2D detections h:**

$$p(H|I) = \sum_h p(H|h,I)p(h|I) \sim \sum_h p(h|H)p(H)p(h|I)$$

3D prior

- **3D prior**
  - **Distance prior (uniform range)**
  - **Size prior (Gaussian)**

$\Rightarrow$ Significantly reduced search space



2D/3D Interactions

- **Likelihood of 3D hypothesis H given image I and 2D detections h:**

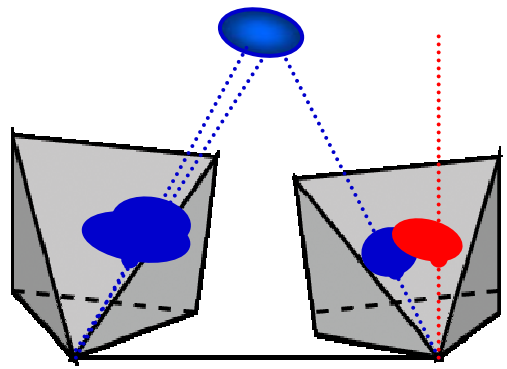
$$p(H|I) = \sum_h p(H|h,I)p(h|I) \sim \sum_h p(h|H)p(H)p(h|I)$$

2D/3D transfer

- **2D/3D transfer**
  - **Two image-plane detections are consistent if they correspond to the same 3D object**

  ⇒ Multi-viewpoint integration

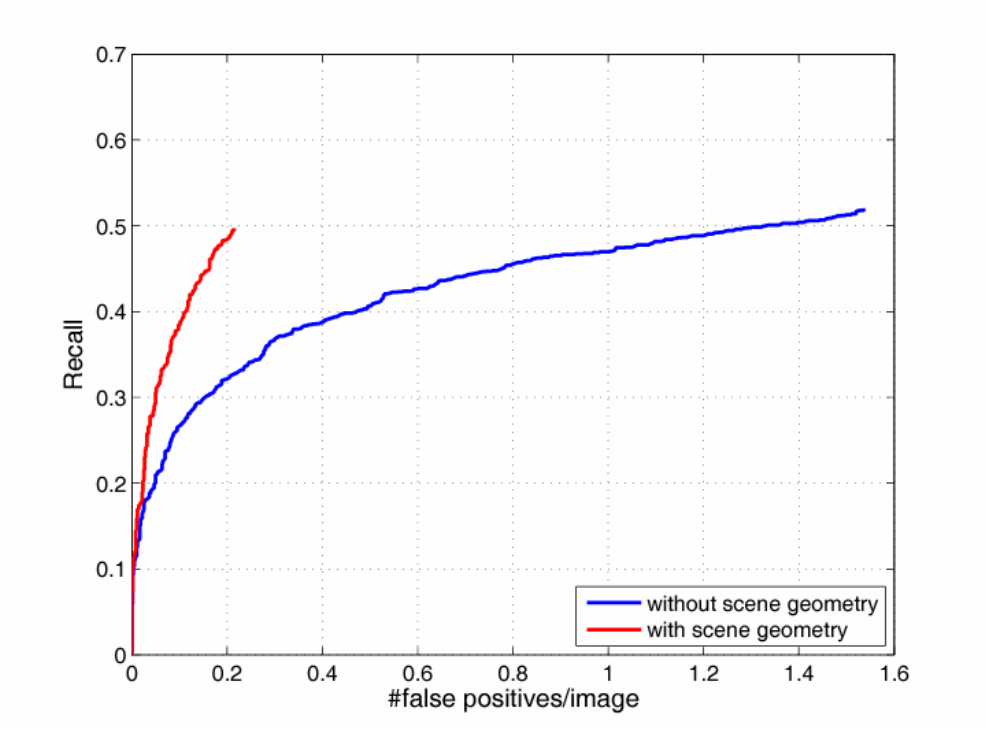
  ⇒ Multi-camera integration



Detections Using Ground Plane Constraints

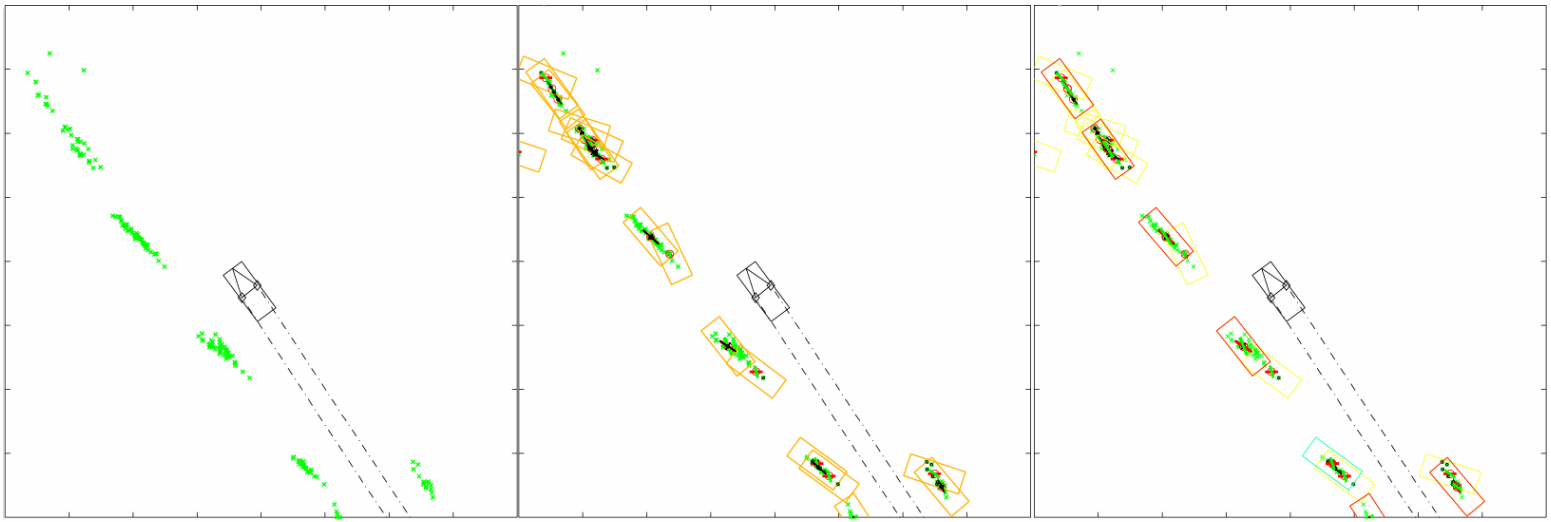
left camera
1175 frames

Quantitative Results



- **Detection performance on first 600 frames**
  - **All cars annotated that were >50% visible**
  - **Ground plane constraint significantly improves precision**
  - **Performance: 0.2 fp/image at 50% recall**

Temporal Integration



- **Temporal integration in world coordinate frame**
  - **Using external camera calibration from SfM.**
  - **Each detection transfers to a 3D observation H.**
  - **Find superset of 3D hypotheses $\mathcal{H}$.**
  - **Estimate orientation using cluster shape & detected viewpoints.**
  - **Select set of 3D hypotheses that best explain the observations.**

Hypothesis Selection for 3D Detections

- **Quadratic Boolean Optimization Problem (from MDL)**

$$\max_m m^T Q m = m^T \begin{bmatrix} q_{11} & \cdots & q_{1M} \\ \vdots & \ddots & \vdots \\ q_{M1} & \cdots & q_{MM} \end{bmatrix} m$$

[Leonardis et al,95]

- **Individual scores (diagonal terms)**

$$q_{ii} = -\tilde{\kappa}_1 + \sum_{H \in \mathcal{H}_i} e^{-(t-t_i)/\tau} \left((1-\tilde{\kappa}_2) + \tilde{\kappa}_2 \, p(H|\mathcal{H}_i) p(H|I) \right)$$

- **Interaction costs (off-diagonal terms)**

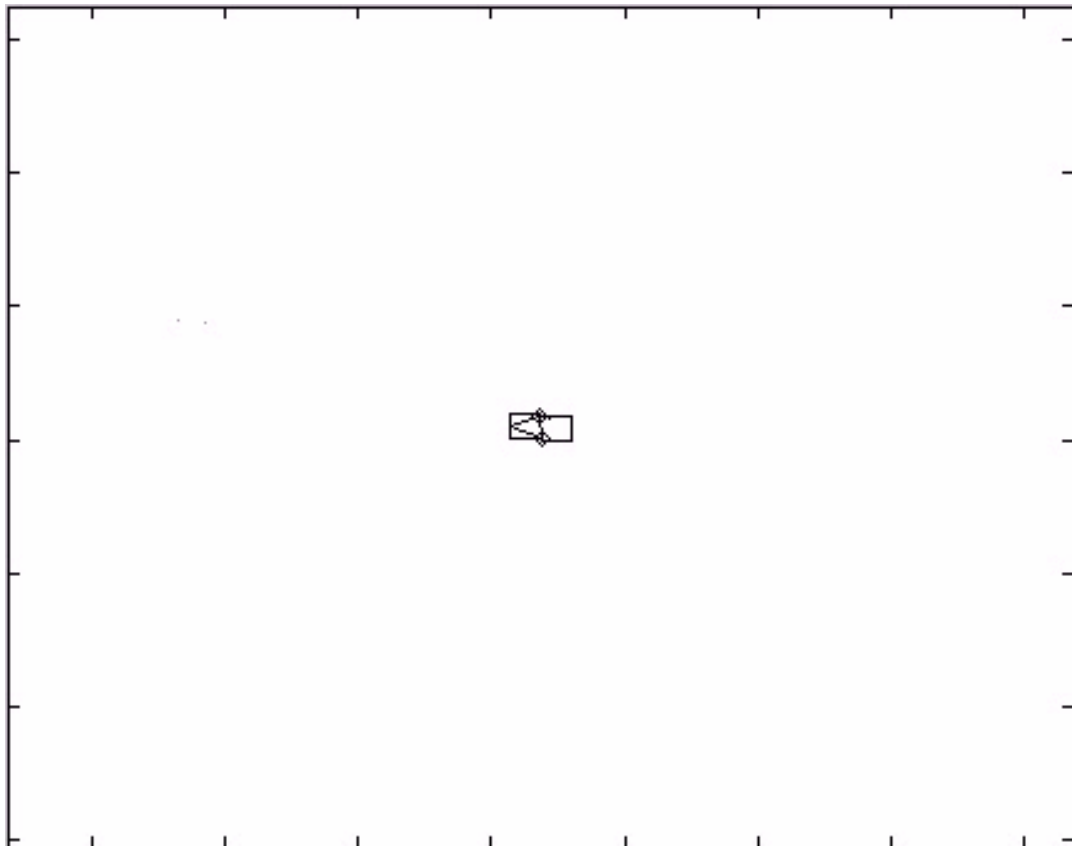
$$q_{ij} = -\frac{1}{2} \sum_{H \in \mathcal{H}_i \cap \mathcal{H}_j} e^{-(t-t^*)/\tau} \left((1-\tilde{\kappa}_2) + \tilde{\kappa}_2 \, p(H|\mathcal{H}^*) p(H|I) + \tilde{\kappa}_3 \, O(\mathcal{H}_i, \mathcal{H}_j) \right)$$

temporal decay

likelihood of membership to hypothesis $\mathcal{H}$

penalty for physical overlap

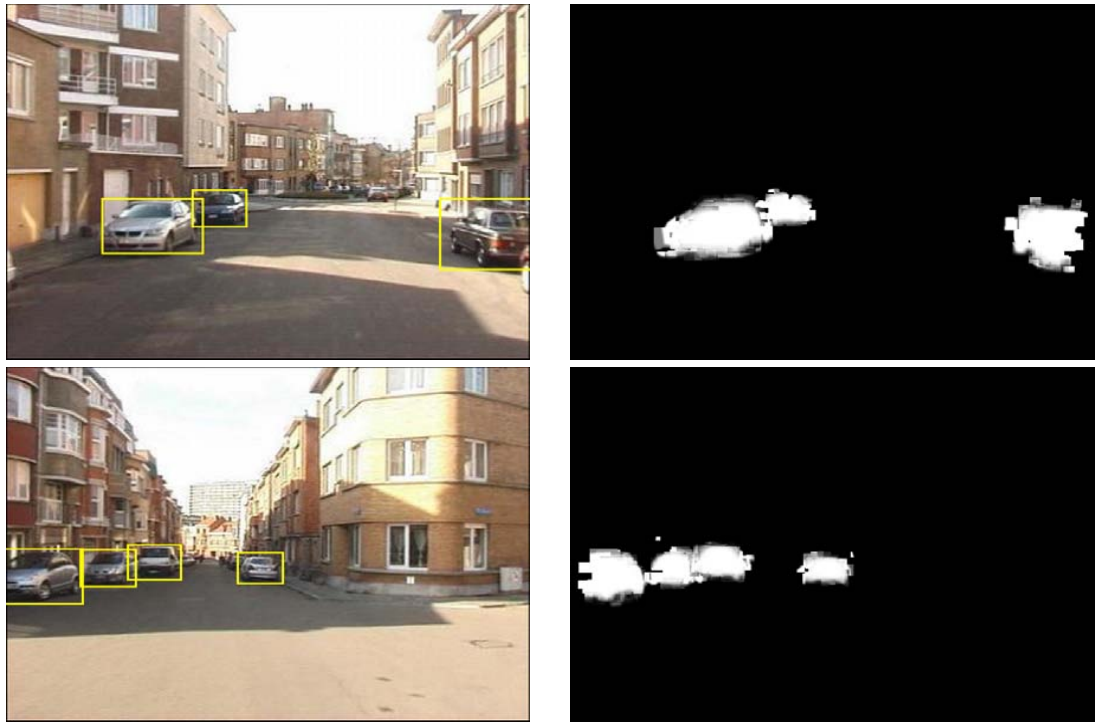
Result of Temporal Integration



Online 3D Car Location Estimates

3D Estimates After Convergence

Feedback into 3D Reconstruction



- **Feedback of detections & segmentation maps**
  - **Used to discard features on cars for SfM**
  - **Used to mask out cars in dense reconstruction**

  ⇒ **More accurate 3D estimates** *in the next frame.*

Another Application: 3D City Modeling

Enhancing your driving experience...

Original

3D Reconstruction

Conclusion

- **Object detection in crowded scenes**
  - Cognitive Loop between recognition and segmentation
  - Initial hypotheses → top-down segmentation → verification
- **Application for cognitive scene analysis**
  - Combining recognition, reconstruction, and temporal integration
- **Cognitive Loop between 2D and 3D processing**
  - Reconstruction delivers camera calibration, ground plane
  - 3D context tremendously improves recognition performance
  - Car detection, segmentation makes 3D estimation more accurate
- **System applied to challenging real-world task**
  - Real-time 3D reconstruction (26-30 fps)
  - Accurate object detection & 3D pose estimation results

Thank you very much for your attention!