Recent Advances in Pedestrian Detection

Bernt Schiele
TU Darmstadt

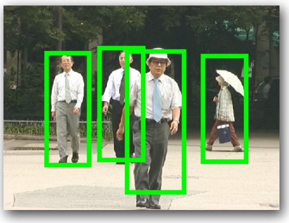
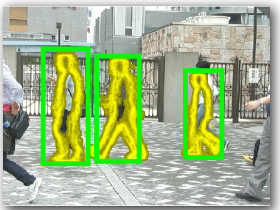
joint work with: Edgar Seemann, Bastian Leibe, Krystian Mikolajczyk, Mario Fritz

Multimodale interaktive Systeme

Pedestrian Detection in Realistic Environments

Multimodale interaktive Systeme

- General Object Detection Challenges:
  - clutter, partial occlusion, illumination, ...
- For Pedestrians:
  - body articulation greatly influences appearance
- Fundamental Ideas:
  - learn and recognize possible body articulations
  - explicitly share features across body articulations

Overview

- Implicit Shape Model (ISM)
  - [Leibe, Seemann, Mikolajczyk, Schiele cvpr05, bmvc05]
- 4D-Implicit Shape Model (4D-ISM)
  - [Seemann, Leibe, Schiele cvpr06]
- Cross-Articulation Learning ⇨ Explicit Feature Sharing
  - [Seemann, Schiele dagm06]
- SVM-Verification
  - [Seemann, Fritz, Schiele]

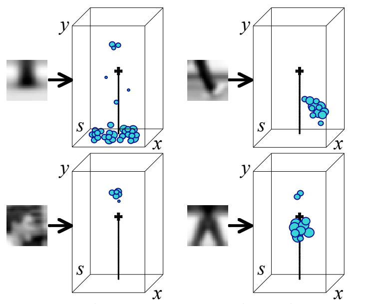
Bernt Schiele - TU Darmstadt

Implicit Shape Model (ISM) - Representation

Training images + segmentation mask

Appearance Codebook

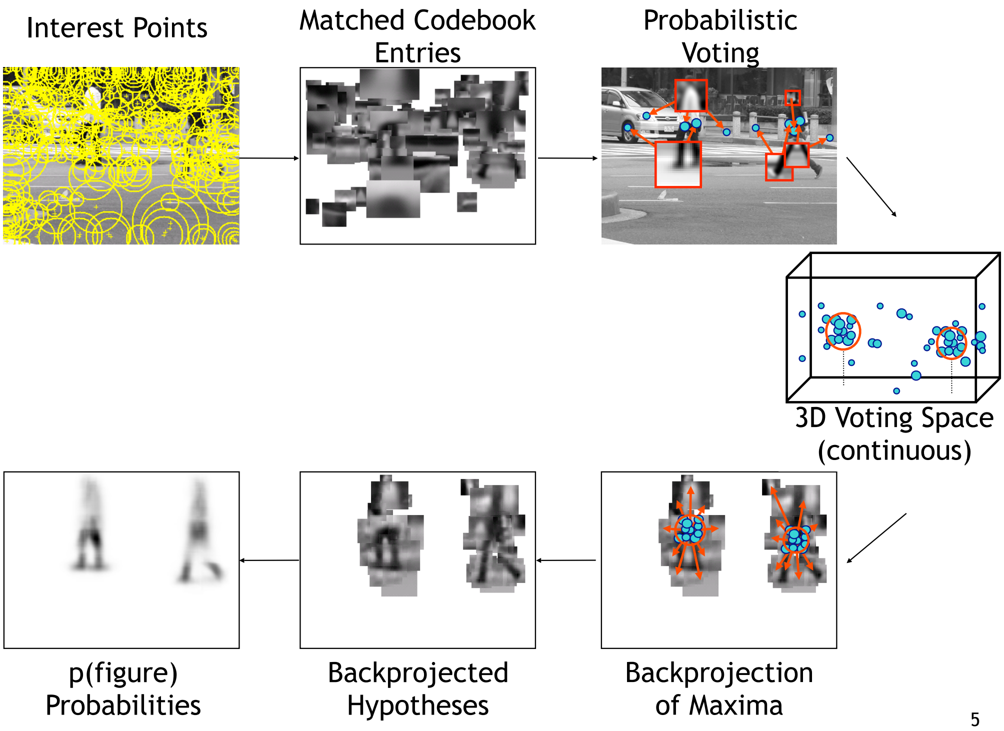
- Learn Appearance Codebook
  - extract features at DoG interest points
  - agglomerative clustering ⇒ codebook
- Learn Spatial Occurrence Distributions
  - match codebook to training images
  - record 3D-distributions: location(x,y) & scale



Spatial occurrence distributions

Bernt Schiele - TU Darmstadt

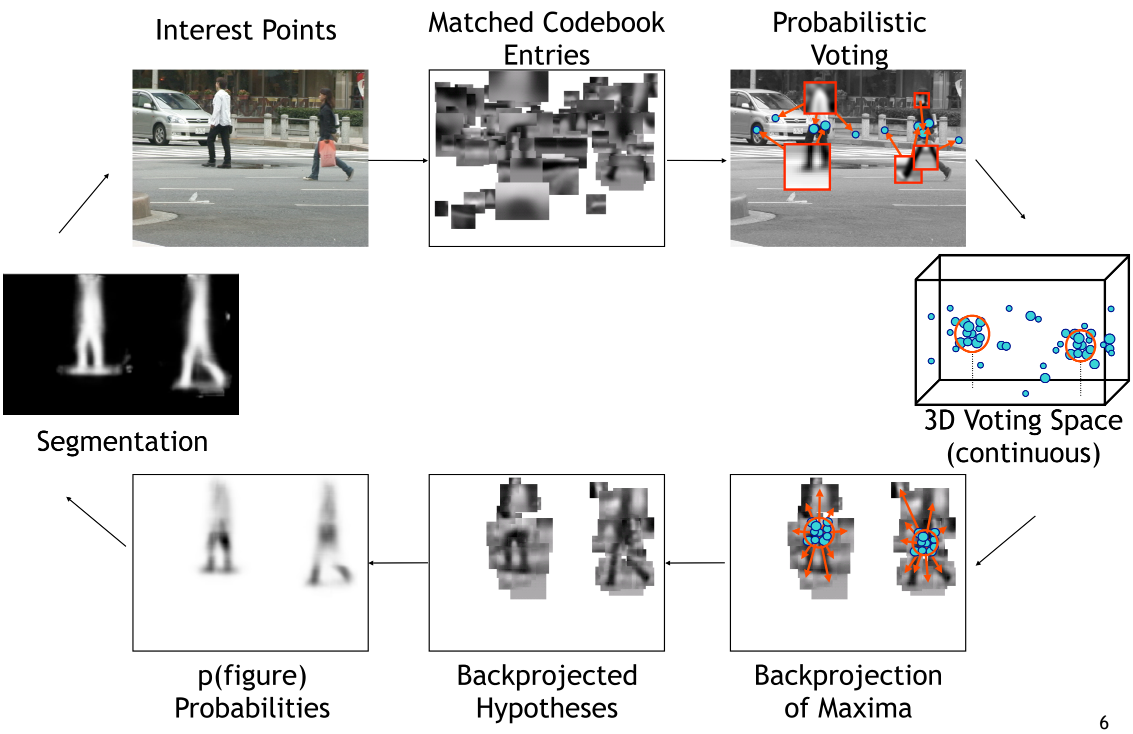
Implicit Shape Model (ISM) - Recognition



Bernt Schiele - TU Darmstadt

Implicit Shape Model (ISM) - Recognition

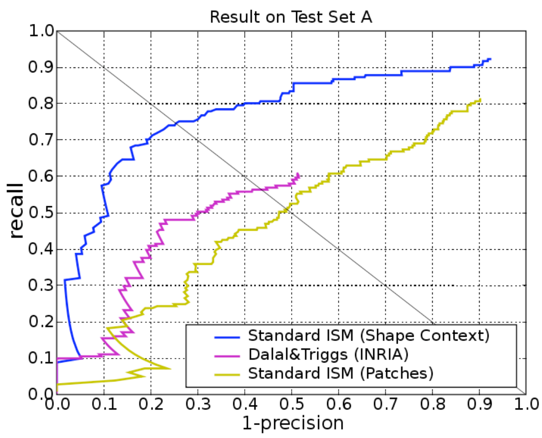
Multimodale interaktive Systeme

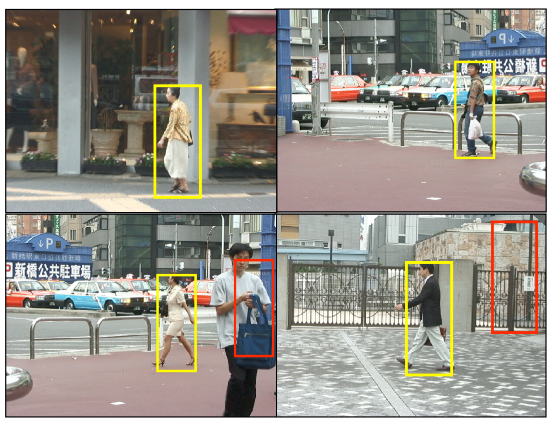


Experimental Setup

- Training:
  - 210 side views
  - two backgrounds
- Test Set A
  - 181 street scene images
  - one pedestrian per image
- Test Set B - ‘crowded scenes’
  - 209 street scene images
  - 595 pedestrians in total

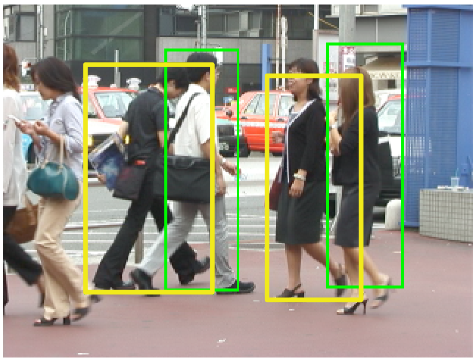
Results - Standard Implicit Shape Model

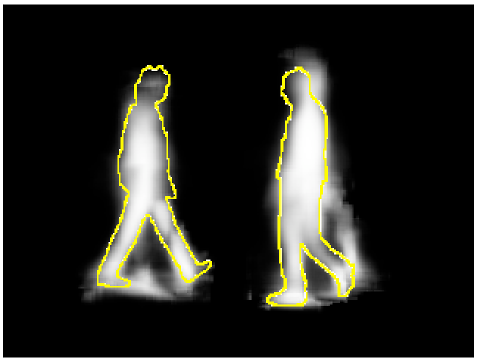




- Good performance, when using shape context as feature
- Competitive w.r.t. other state-of-the-art methods (ISM trained on side-views only)

Problem for Articulated Objects





- Over-complete Segmentations
  - flexible spatial model
  - segmentations may contain superfluous body parts
    ⇨ score of neighboring hypotheses may be reduced!
- Idea: Enforce Global Consistency
  - silhouette verification [cvpr05]
  - 4D-ISM [cvpr06]

Overview

- Implicit Shape Model (ISM)
  - [Leibe, Seemann, Mikolajczyk, Schiele cvpr05, bmvc05]
- 4D-Implicit Shape Model (4D-ISM)
  - [Seemann, Leibe, Schiele cvpr06]
- Cross-Articulation Learning ⇨ Explicit Feature Sharing
  - [Seemann, Schiele dagm06]
- SVM-Verification
  - [Seemann, Fritz, Schiele]

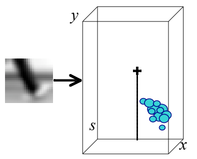
4D-ISM

- Learn typical articulations by silhouettes clustering:

Fig.: Resulting articulation clusters

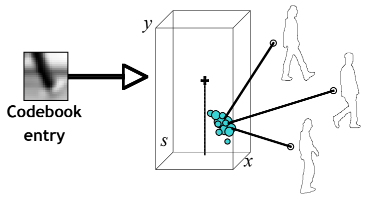
- Learning the occurrence distribution:
  - 3D-distribution of feature: location (x,y) & scale
  - +1D: on which articulation cluster the feature occurs (pose)

3D Occurrence Distributions



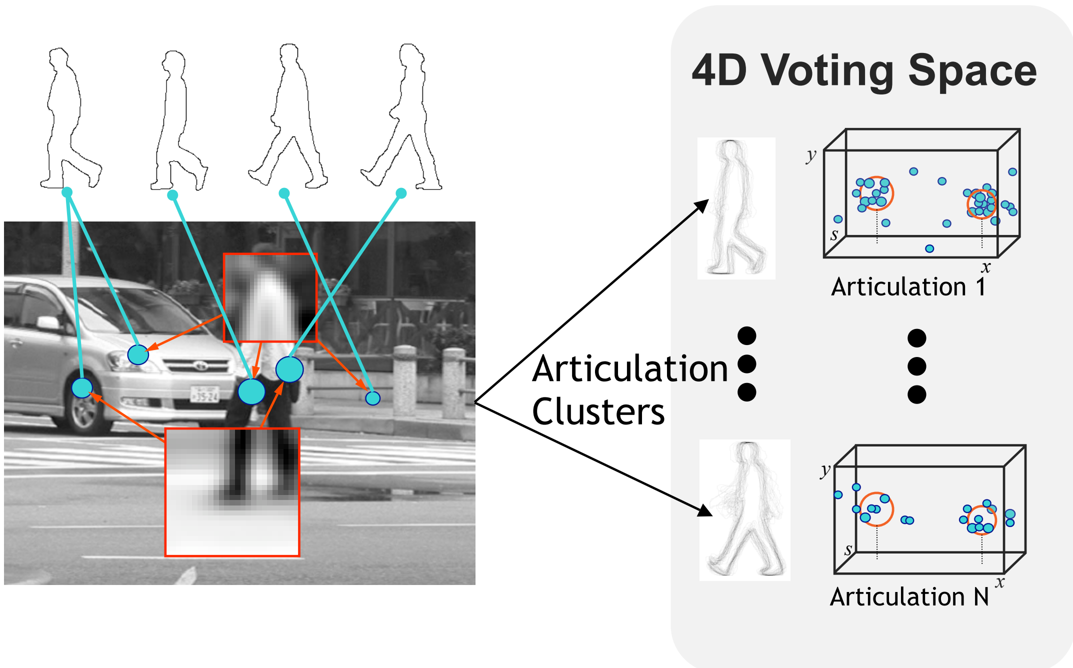
Vote v = (pos_x, pos_y, scale)

4D Occurrence Distributions



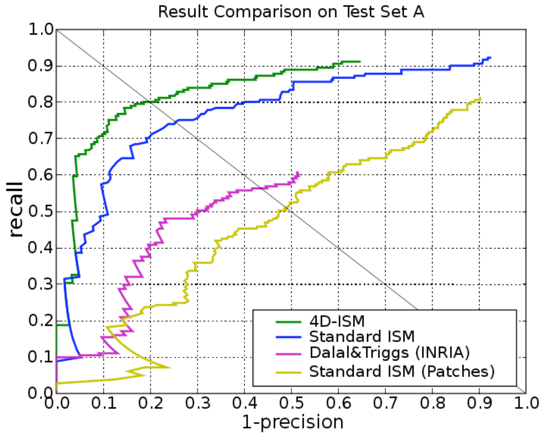
Vote v = (pos_x, pos_y, scale, pose)

4D-ISM - Voting



- Resulting hypotheses are consistent w.r.t. body articulations

Results - 4D-ISM



- 6% improvement in EER
- reduces false positives
- more flexible than global silhouette verification (can handle partial occlusions)

Silhouette verification [cvpr05]

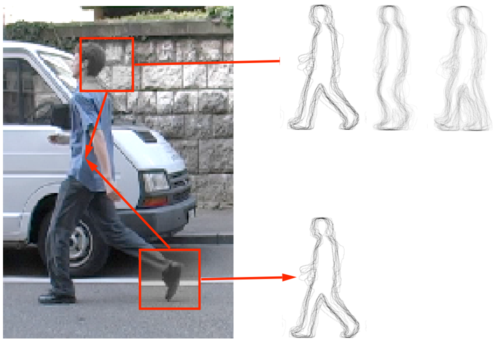
4D-ISM [cvpr06]

Overview

- Implicit Shape Model (ISM)
  - [Leibe, Seemann, Mikolajczyk, Schiele cvpr05, bmvc05]
- 4D-Implicit Shape Model (4D-ISM)
  - [Seemann, Leibe, Schiele cvpr06]
- Cross-Articulation Learning ⇨ Explicit Feature Sharing
  - [Seemann, Schiele dagm06]
- SVM-Verification
  - [Seemann, Fritz, Schiele]

Cross-Articulation Learning

Training image



Idea:

- explicitly share features across articulations
- less training data needed
- better generalization

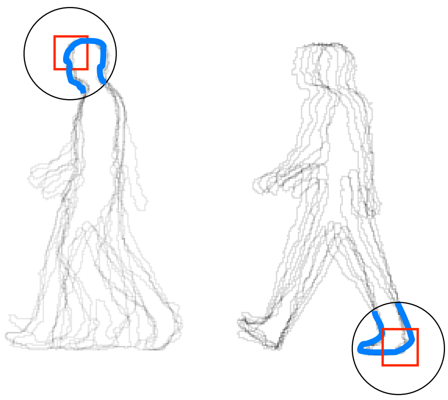
⇨ learn for each feature, with with articulations it is consistent

Explicit Feature Sharing

- For each feature:
  - check consistency with all articulations by matching local contexts / neighborhoods

  ⇨ Share feature if local context is similar

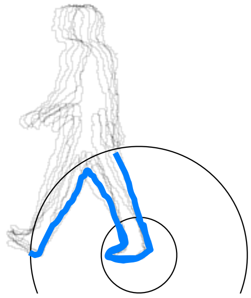
- Local context matching is independent of feature descriptor
  - Local silhouette segments
  - Shape context descriptors

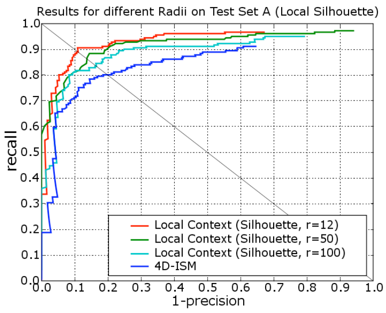


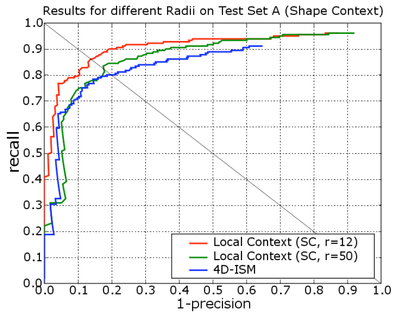
Local Context & Feature Sharing

Multimodale interaktive Systeme

- Local Context Radius
  - can be varied and determines 'locality' of feature sharing
  - 4D-ISM is special case for radius = "object size"
- Smaller Context Radius
  - allows more feature sharing & performs better





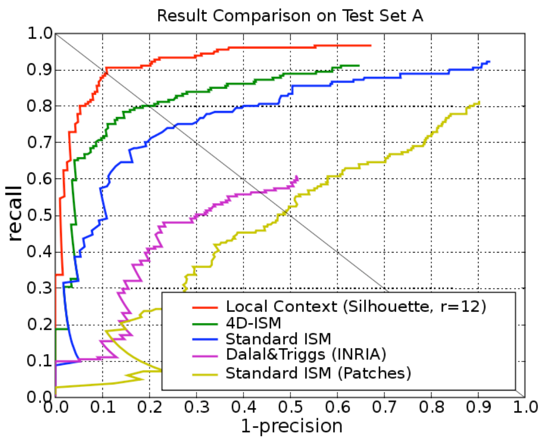


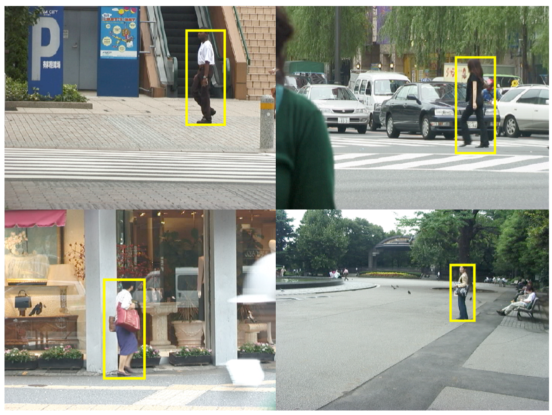
- Cross-Articulation learning from clean silhouettes is superior to local shape context regions (with background)

Bernt Schiele - TU Darmstadt

Comparison to Previous Results (Test Set A)

Multimodale interaktive Systeme



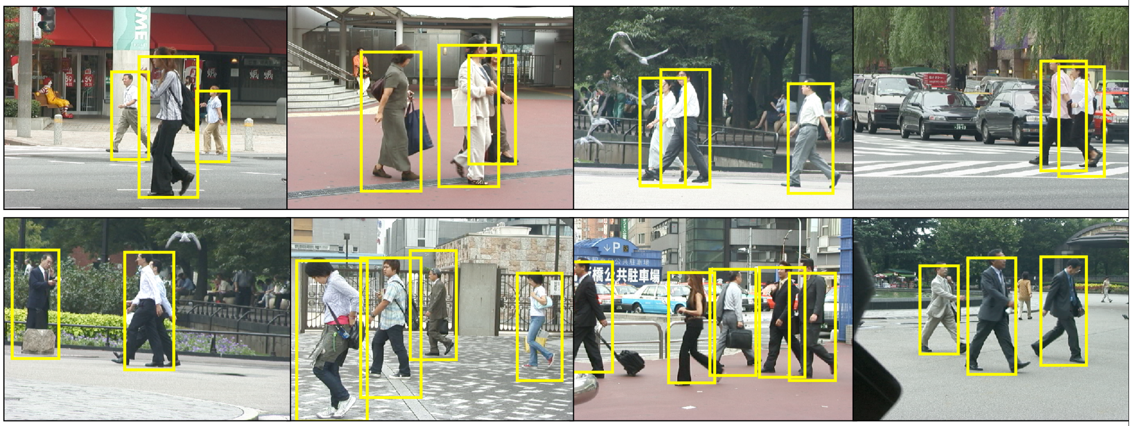


- Using body articulations improves EER by 15%
  - 5% from 4D-ISM
  - 10% from cross-articulation learning
- Cross-Articulation learning from clean silhouettes is superior to local shape context regions (with background)

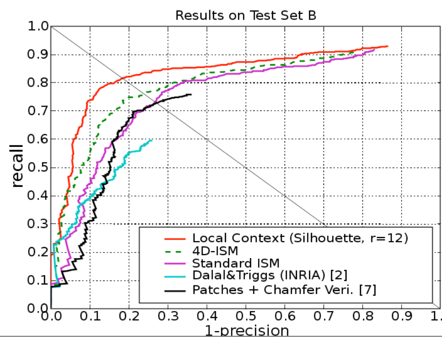
Bernt Schiele - TU Darmstadt

Multimodale interaktive Systeme

Detections on ‘Crowded Scenes’ (Test Set B)



- Use of Articulations improves EER by 8-9%
  - cross-articulation learning 4-5% improvement over 4D-ISM
- Better precision throughout



Bernt Schiele - TU Darmstadt

Multimodale interaktive Systeme

Overview

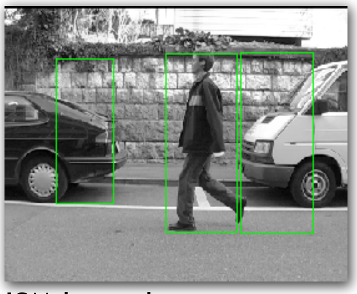
- Implicit Shape Model (ISM)
  - [Leibe, Seemann, Mikolajczyk, Schiele cvpr05, bmvc05]
- 4D-Implicit Shape Model (4D-ISM)
  - [Seemann, Leibe, Schiele cvpr06]
- Cross-Articulation Learning ⇨ Explicit Feature Sharing
  - [Seemann, Schiele dagm06]
- SVM-Verification
  - [Seemann, Fritz, Schiele]

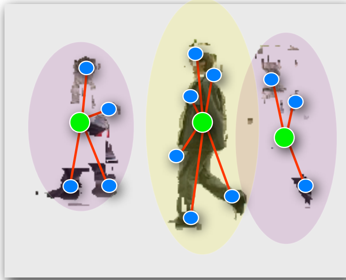
Bernt Schiele - TU Darmstadt

ISM with Integrated SVM Verification

Multimodale interaktive Systeme


Input image


ISM hypotheses


SVM training examples

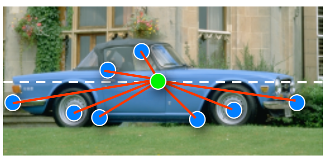
- Learn a discriminative detection model
  - as opposed to the generative nature of the ISM
- Learning on top of ISM output
  - directly use local feature representation of ISM
- SVM verification based on the spatial relationships of local features

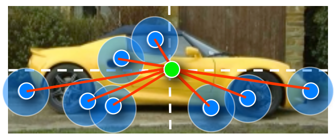
Bernt Schiele - TU Darmstadt

Local Kernel SVM

Multimodale interaktive Systeme

- uses local features similarity kernel:


X


Y

$$K((x,\lambda_x,s_x),(y,\lambda_y,s_y)) = \underbrace{exp(-\gamma(1-d(x,y)))}_{\text{appearance similarity}} \cdot \underbrace{exp(-\frac{\lambda_x-\lambda_y}{2\sigma_\lambda^2})}_{\text{position constraint}} \cdot \underbrace{exp(-\frac{log(s_x-s_y)}{2\sigma_s^2})}_{\text{scale constraint}}$$

Bernt Schiele - TU Darmstadt

Local Kernel SVM

- Kernel to match sets of local features inspired by
  - [Wallraven03,Caputo04,Fritz05]

$$K(X,Y) = \frac{1}{k} \max_{\Phi,\Psi} \sum_{j=1}^{k} K_l((x_{\Phi(j)}, \lambda_{x,\Phi(j)}, s_{x,\Phi(j)}), (y_{\Psi(j)}, \lambda_{y,\Psi(j)}, s_{y,\Psi(j)}))$$

maximum over permutations

local feature similarity kernel

descriptor

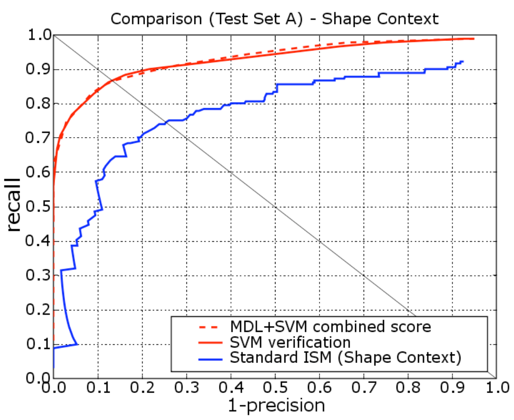
position

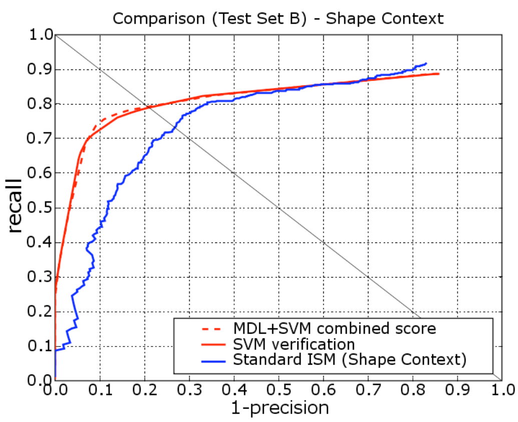
scale

- greedy approximation of maximum/matching
- non-mercer kernel
  - in most practical settings kernel matrix is positive definite [Boughorbel04]

Bernt Schiele - TU Darmstadt

ISM & SVM



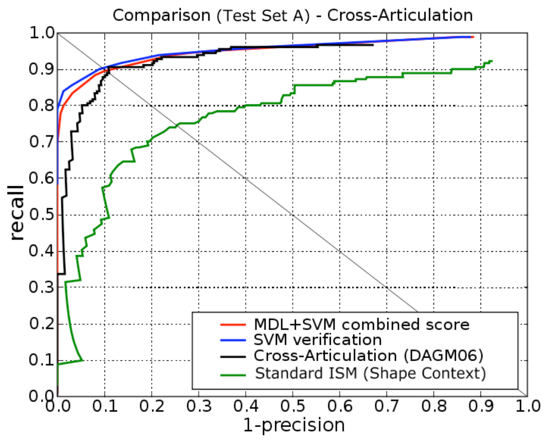


- SVM improves EER by 11%
- Precision considerably better at 70-80% recall

- 7% improvement in EER for overlapping pedestrians

Bernt Schiele - TU Darmstadt

Cross-Articulation Learning & SVM



- Training SVM on top of cross-articulation learning
  - further improves performance
  - Detection precision is particularly increased

- Overall Improvement
  - 15% EER through by using explicitly articulation clusters
  - 5-10% EER through the use of cross-articulation learning
  - SVM increased precision

‘Crowded Scene’ Movie

- single frame detection (no temporal information used)
  - yellow = true positive detections
  - red = false positive detections