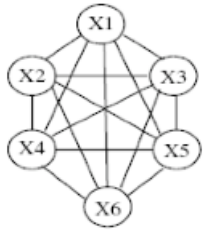
Part-Based Object and People Detection
Cognitive Science Summerschool 2oo9

TECHNISCHE UNIVERSITÄT DARMSTADT

Part 3: Global Features Models

e.g. HOG: fixed spatial relationships



Bernt Schiele

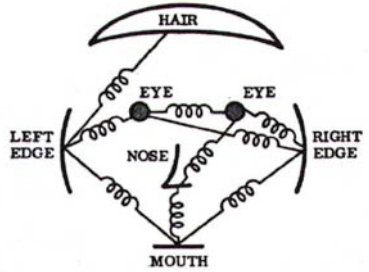
Department of Computer Science
TU Darmstadt, Germany

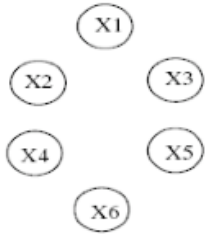


Overview

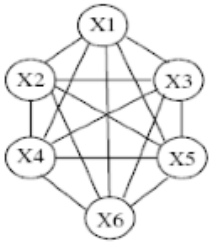
- Introduction (part 1)
  - why study computer vision in general and object recognition in particular :)

- Object Recognition Methods
  - Bag of Words Models (BoW) (part 2)
    - Model: Histogram of local features
    - e.g. Interest Points (scale invariant)
  - Global Feature Models + Classifier (part 3)
    - e.g. HOG = Histogram of Oriented Gradients
      - global object feature / description
    - e.g. SVM = Support Vector Machines
      - discriminant classifier - widely used
  - Part-Based Object Models (part 4)
    - e.g. Implicit Shape Model (ISM)
    - local parts & global constellation of parts

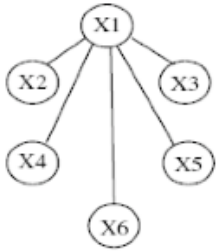




BoW: no spatial relationships



e.g. HOG: fixed spatial relationships



e.g. ISM: flexible spatial relationships

Global Feature Models
Example: static HOG-feature vector

TECHNISCHE UNIVERSITÄT DARMSTADT

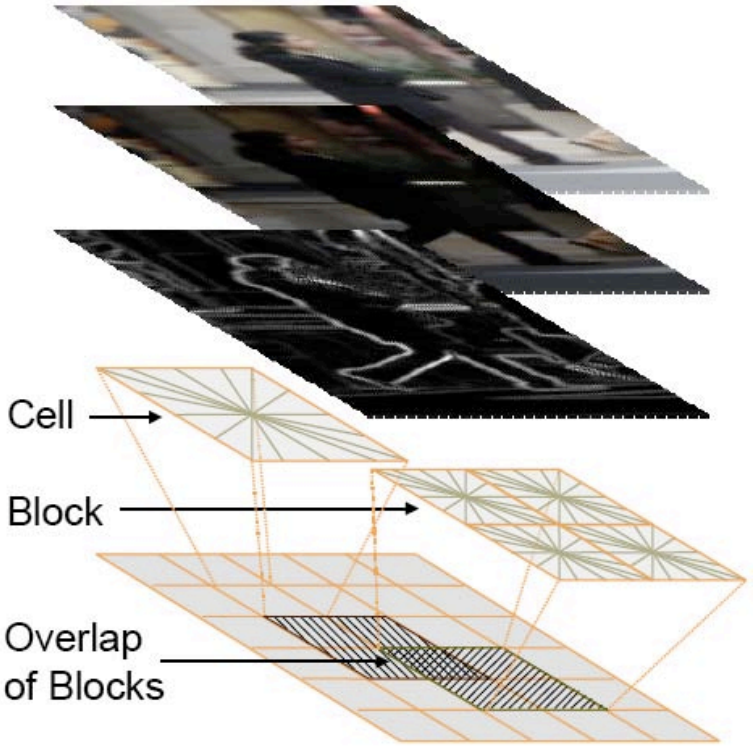


Input Image

Detection Window

- Normalise gamma
- Compute gradients
- Weighted vote in spatial & orientation cells
- Contrast normalise over overlapping spatial cells
- Collect HOGs over detection window
- Linear SVM

HOG: Histogram of Oriented Gradients



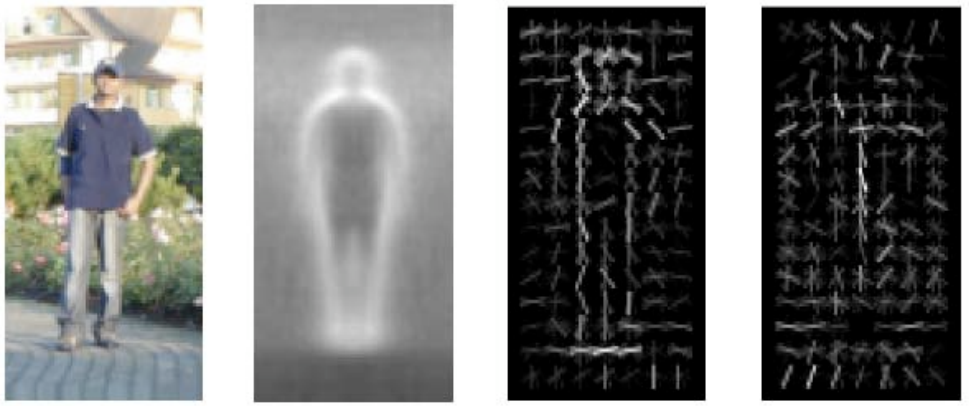
Feature vector f = […, …, …]

[Dalal,Triggs@CVPR-05]

Global Feature Models
Example: static HOG-feature vector

TECHNISCHE UNIVERSITÄT DARMSTADT

- Most Important Cues:
  - head, shoulder, leg silhouettes
  - vertical gradients inside a person are counted as negative
  - overlapping blocks just outside the contour are most important



Input example | Average gradients | Weighted pos wts | Weighted neg wts

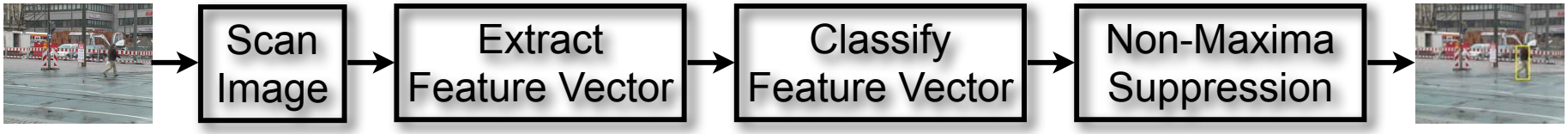
- "Local Context" Use:
  - Note that Dalal & Triggs obtain best performance by including quite substantial context/background around the person

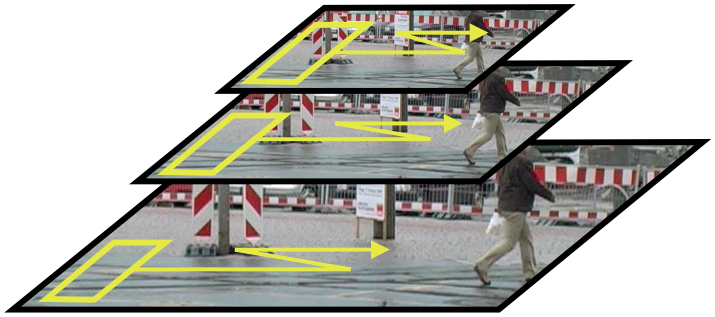
[Dalal,Triggs@CVPR-05]

Global Feature Models: Sliding Window Method for Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- Sliding Window Based Object & People Detection:





‘slide’ detection window
over all positions & scales

Two Important Questions:
1) which feature vector
2) which classifier

Overview of lecture parts 3 & 4...

TECHNISCHE UNIVERSITÄT DARMSTADT

- Global Feature Based Methods for People Detection (part 3)
  - A Performance Evaluation of Single and Multi-Feature People Detection [Wojek,Schiele@DAGM-08]
  - Pedestrian Detection: A New Benchmark [Dollar,Wojek,Perona,Schiele@CVPR-09]
  - Multi-Cue Onboard Pedestrian Detection [Wojek,Walk,Schiele@CVPR-09]
- Part-Based Model for People & Object Detection (part 4)
  - Detection by Tracking and Tracking by Detection [Andriluka,Roth,Schiele@CVPR-08]
  - Pictorial Structures Revisited: People Detection and Articulated Pose Estimation [Andriluka,Roth,Schiele@CVPR-09]
  - A Shape-Based Object Class Model for Knowledge Transfer [Stark,Goesele,Schiele@ICCV-09]

Comparison of Pedestrian Detectors

TECHNISCHE UNIVERSITÄT DARMSTADT

- Motivation (global feature & sliding window approaches):
  - what is the best feature?
  - what is the best classifier?
  - how complementary are those features/classifier?

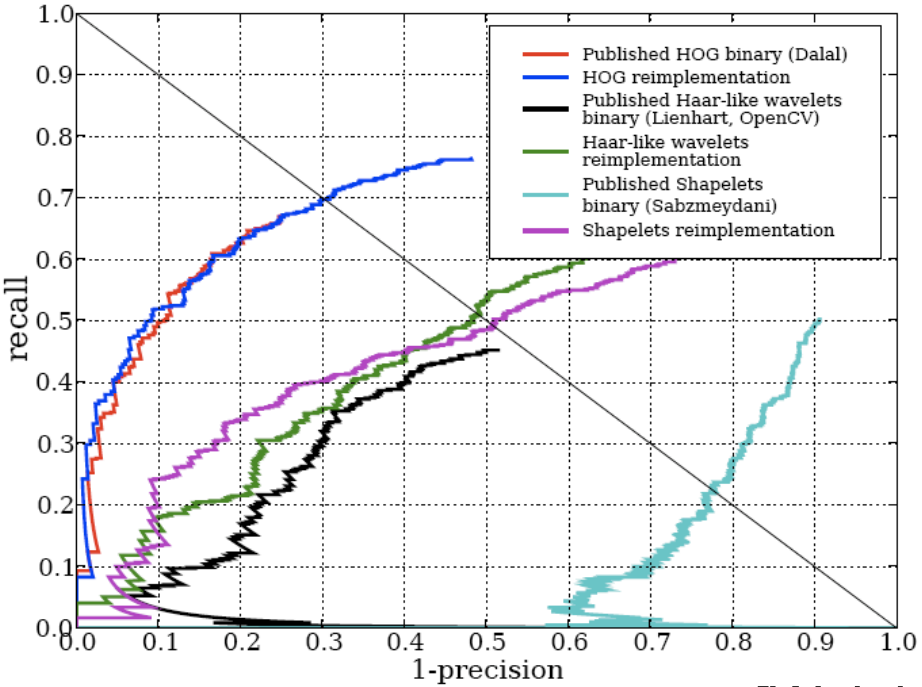
- Overview (sliding window approaches):

	Linear SVM	Kernel SVM	Ada-Boost	Other	Criterion
Haar wavelet [1]		poly-nomial			ROC
Haar-like wavelet [10]			cascaded		ROC
HOG [2]	✓	RBF			FPPW
Shapelets [13]			✓		FPPW
Shape Context [8]				ISM	RPC

[Wojek,Schiele@DAGM-08]

Reimplementation of Approaches

- Comparison of our reimplementations & available binaries
  - our HOG = published binary
  - our Haar Wavelets > openCV-implementation
  - our Shapelets >> published binary

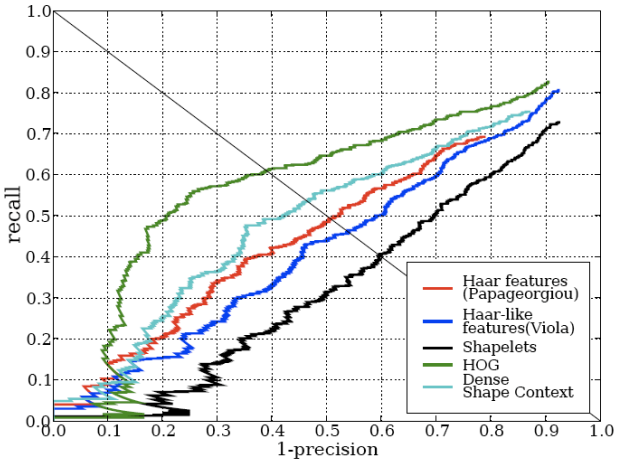


[Wojek,Schiele@DAGM-08]

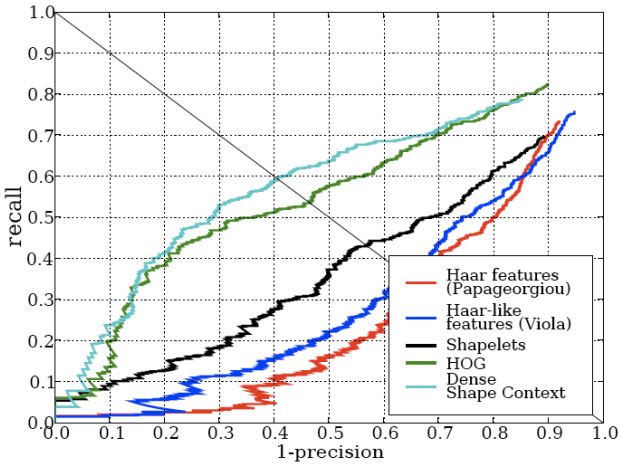
Comparison of different features & classifiers:

TECHNISCHE UNIVERSITÄT DARMSTADT

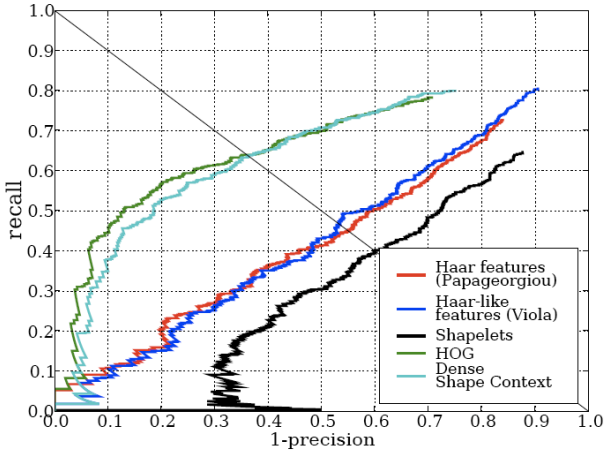
- Same Classifier, different features:



(a) Feature performance with AdaBoost



(b) Feature performance with linear SVM



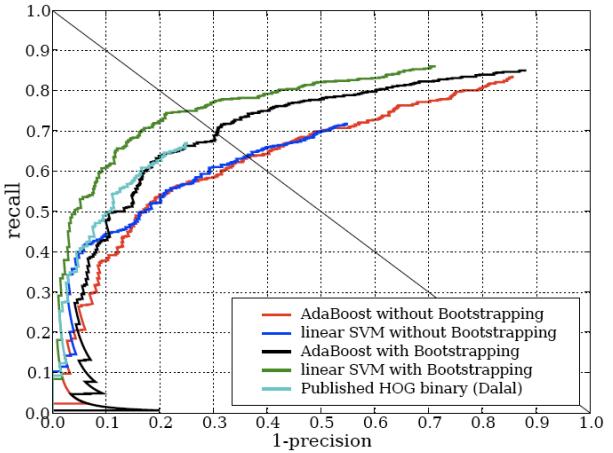
(c) Feature performance with RBF SVM

- Conclusions:
  - best features: HOG & Dense-Shape-Context
  - best classifier: RBF-SVM slightly better than linear SVM

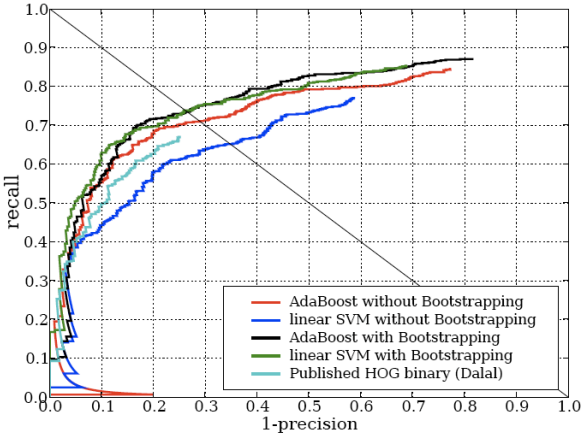
[Wojek,Schiele@DAGM-08]

Combination of different features

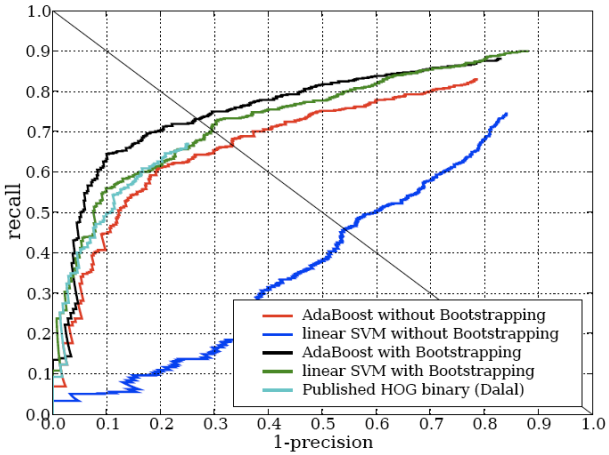
- Combination of
  - Dense Shape Context & Haar Wavelets
  - HOG & Haar Wavelets
  - all three



(e) Combination of Haar wavelets [1] and dense Shape Context, different classifiers



(f) Combination Haar wavelets [1], HOG and dense Shape Context, different classifiers



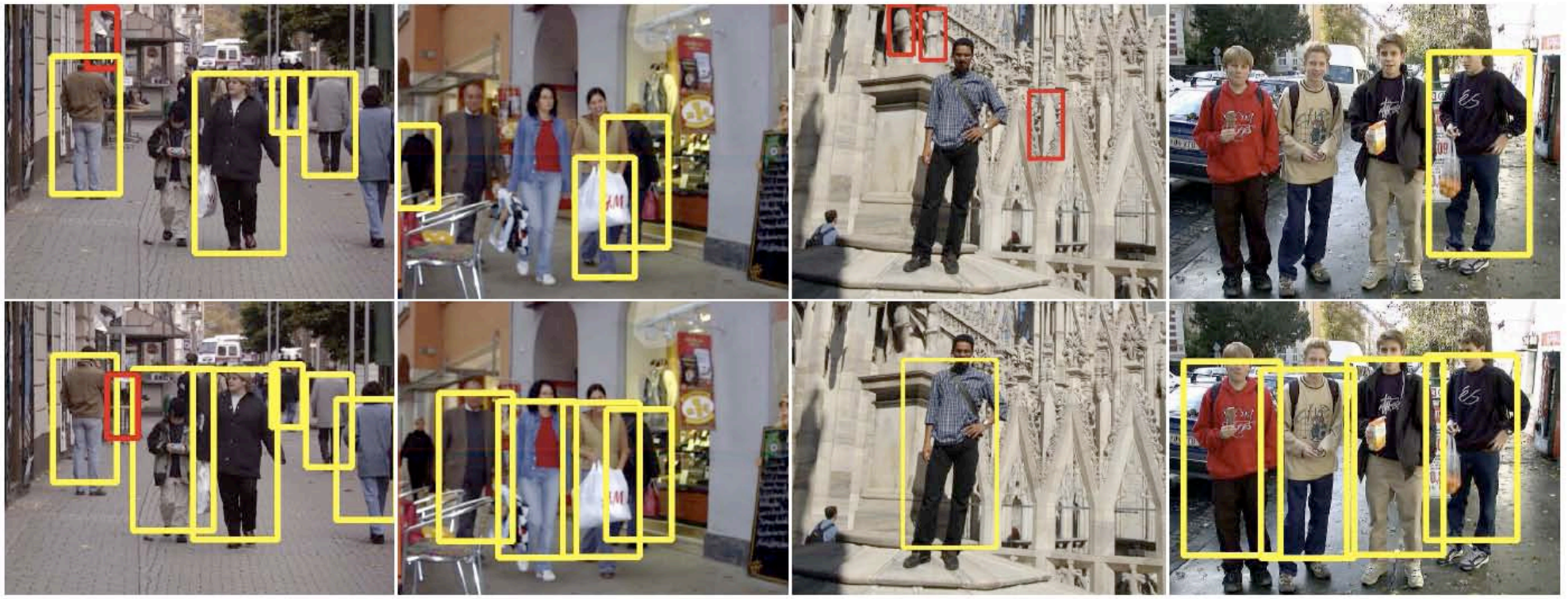
(d) Combination of Haar wavelets [1] and HOG, different classifiers

[Wojek,Schiele@DAGM-08]

Sample Detections: HOG vs. Multi-Feature Combination

TECHNISCHE UNIVERSITÄT DARMSTADT

- Comparison (on INRIA people dataset): [Wojek,Schiele@DAGM-08]
  - 1. row: HOG
  - 2. row: combination of Dense Shape Context & Haar Wavelets



Pedestrian Detection: A New Benckmark

TECHNISCHE UNIVERSITÄT DARMSTADT

- Features of the new Pedestrian Dataset:
  - 11h of ‘normal’ driving in urban environment (greater LA area)
  - annotation:
    - 250’000 frames (~137 min) annotated with 350’000 labeled bounding boxes of 2’300 unique pedestrians
    - occlusion annotation: 2 bounding boxes for entire pedestrian & visible region
    - difference between ‘single person’ and ‘groups of people’





[Dollar,Wojek,Perona,Schiele@CVPR-09]

Comparison to Existing Datasets

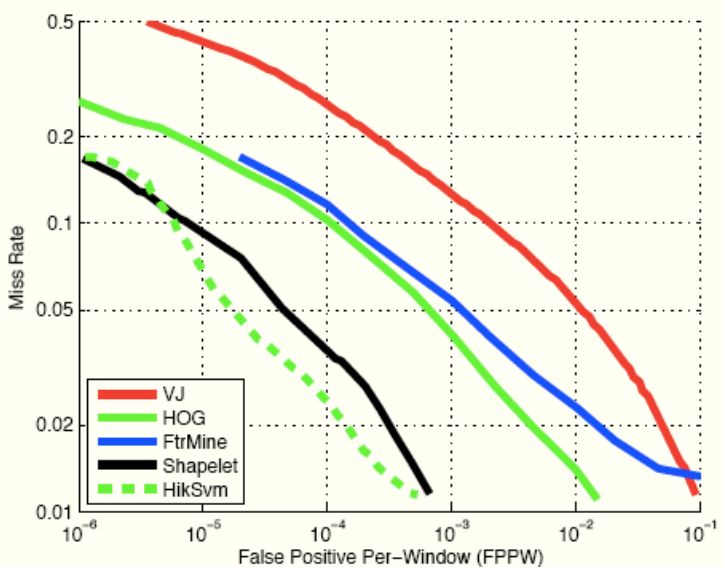
TECHNISCHE UNIVERSITÄT DARMSTADT

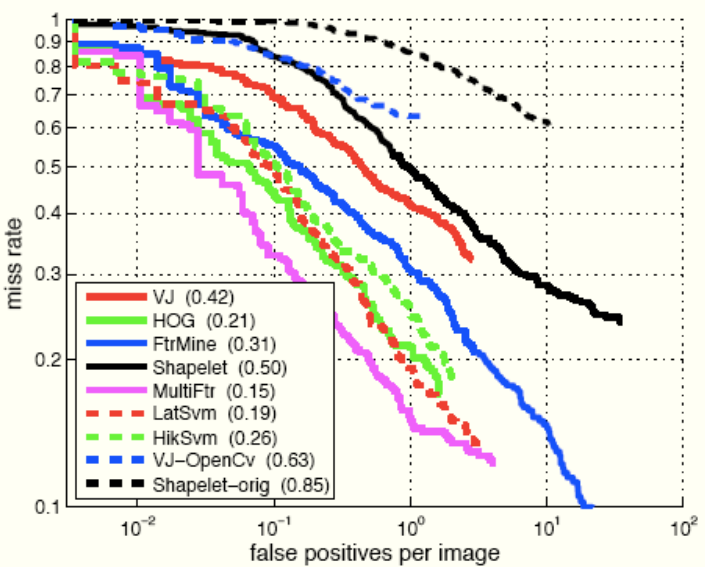
- New Pedestrian Dataset [Dollar,Wojek,Perona,Schiele@CVPR-09]
  - 1 - 2 orders of magnitude larger than any existing dataset
  - new features: temporal correlation of pedestrians, occlusion labeling, ...

	Training			Testing			Height			Properties					
	# pedestrians	# neg. images	# pos. images	# pedestrians	# neg. images	# pos. images	10% quantile	median	90% quantile	color images	per-image ev.	no selec. bias	video seqs.	temporal corr.	occ. labels
MIT[25]	924	–	–	–	–	–	128	128	128	✓					
USC-A[38]	–	–	–	313	–	205	70	98	133		✓				
USC-B[38]	–	–	–	271	–	54	63	90	126		✓				
USC-C[39]	–	–	–	232	–	100	74	108	145		✓	✓			
CVC[13]	1000	6175†	–	–	–	–	46	83	164	✓		✓			
TUD-det[1]	400	–	400	311	–	250	133	218	278	✓	✓				
INRIA[4]	1208	1218	614	566	453	288	139	279	456	✓					
DC[23]	2.4k	15k†	–	1.6k	10k†	–	36	36	36			✓			
ETH[7]	2388	–	499	12k	–	1804	50	90	189	✓	✓	✓	✓		
Proposed	192k	61k	67k	155k	56k	65k	27	48	97	✓	✓	✓	✓	✓	✓

Evaluation Criteria

- Different Approaches:
  - False Positive Per Window (FFPW) vs.
  - False Positives Per Image (FFPI)
- comparison of different algorithms on INRIA-person:
  - ordering of algorithms largely differs between FFPW and FFPI !

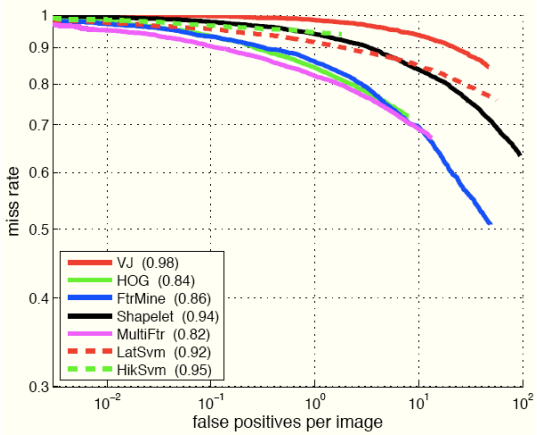




Comparison of Algorithms

- 7 Algorithms tested (FPPI: False-Positives-per-Image):

overall performance

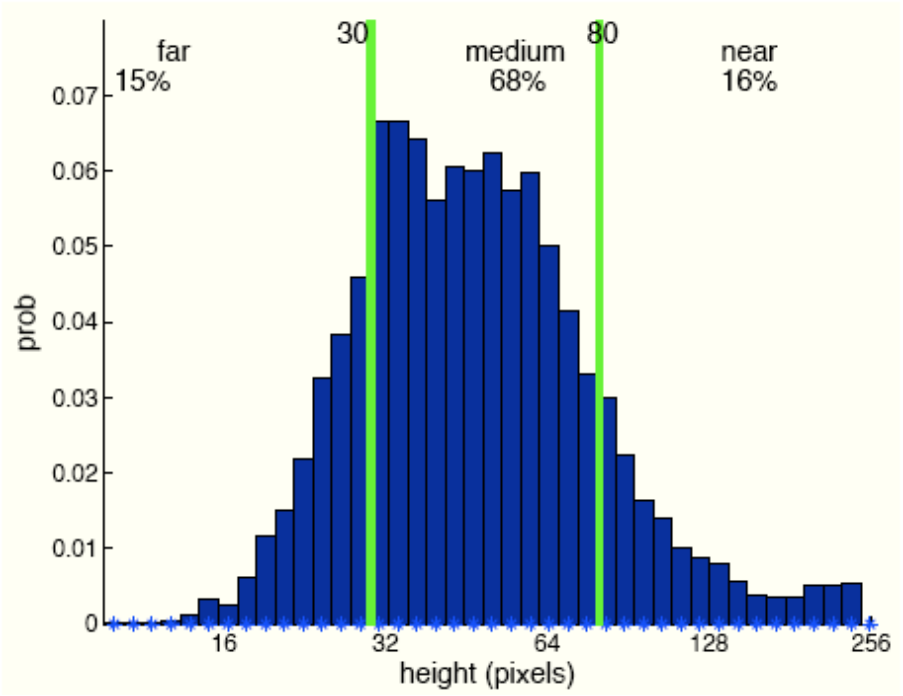


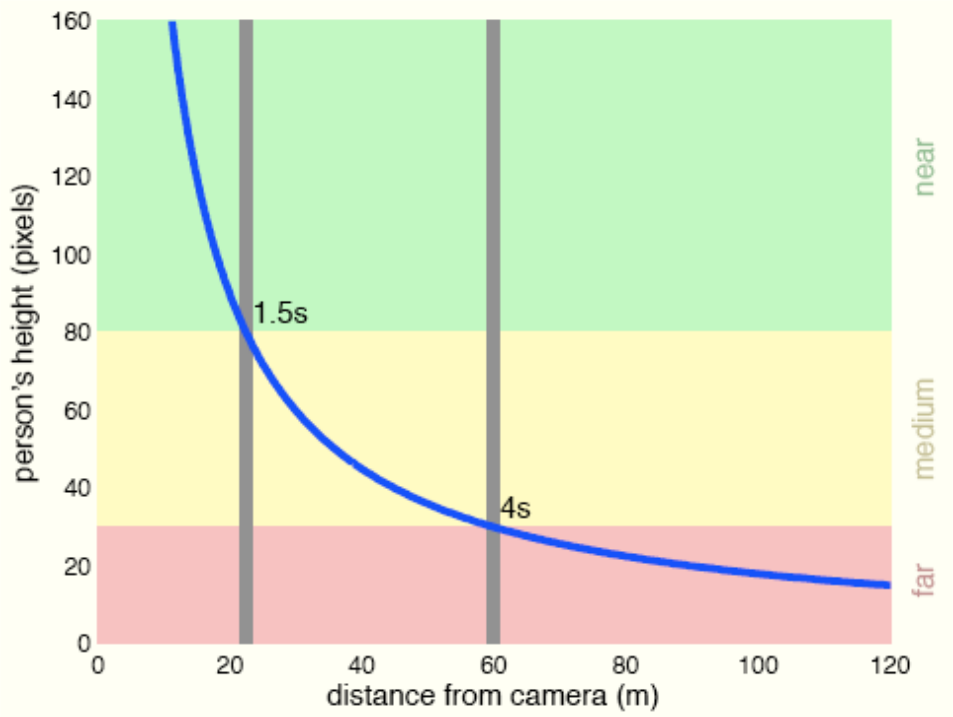
	low-level features	classifier	original implementation	trained on INRIA data	per-image evaluation	sec / frame	model height (in pixels)	scale stride	publication
VJ[35]	Haar	AdaBoost		✓		7.0	100	1.05	'04
HOG[4]	HOG	linear SVM	✓	✓		13.3	100	1.05	'05
FtrMine[6]	gen. Haar	AdaBoost	✓	✓		45	100	1.20	'07
Shapelet[28]	gradients	AdaBoost		✓		60.1	100	1.05	'07
MultiFtr[37]	HOG+Haar	AdaBoost	✓	✓	✓	18.9	100	1.05	'08
LatSvm[10]	HOG	latent SVM	✓		✓	6.3	80	1.05	'08
HikSvm[20]	HOG-like	HIK SVM	✓	✓		140	100	1.20	'08

[Dollar,Wojek,Perona,Schiele@CVPR-09]

Distribution of Pedestrian Sizes

- Differentiation between different sizes:
  - far : pedestrians < 30 pixels large
  - medium : 30 - 80 pixels large
  - near : > 80 pixels large

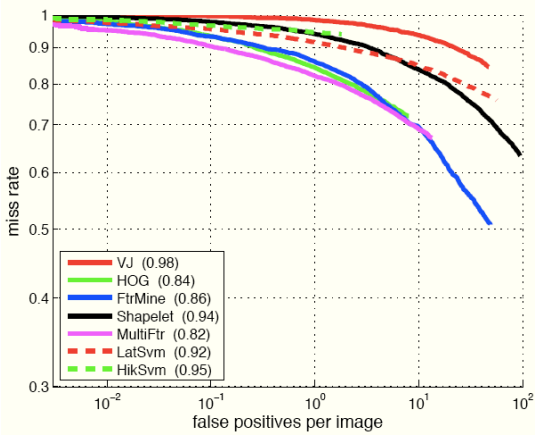




Comparison of Algorithms

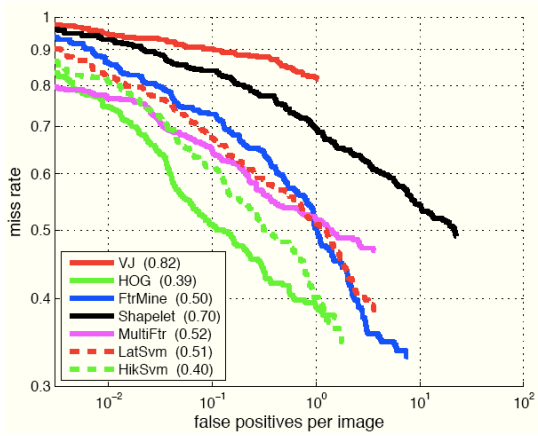
- 7 Algorithms tested (FPPI: False-Positives-per-Image):

overall performance

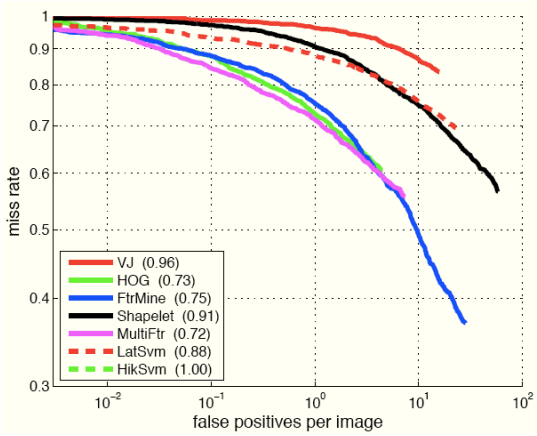


	low-level features	classifier	original im-plementation	trained on INRIA data	per-image evaluation	sec / frame	model height (in pixels)	scale stride	publication
VJ[35]	Haar	AdaBoost		✓		7.0	100	1.05	'04
HOG[4]	HOG	linear SVM	✓	✓		13.3	100	1.05	'05
FtrMine[6]	gen. Haar	AdaBoost	✓	✓		45	100	1.20	'07
Shapelet[28]	gradients	AdaBoost		✓		60.1	100	1.05	'07
MultiFtr[37]	HOG+Haar	AdaBoost	✓	✓	✓	18.9	100	1.05	'08
LatSvm[10]	HOG	latent SVM	✓		✓	6.3	80	1.05	'08
HikSvm[20]	HOG-like	HIK SVM	✓	✓		140	100	1.20	'08

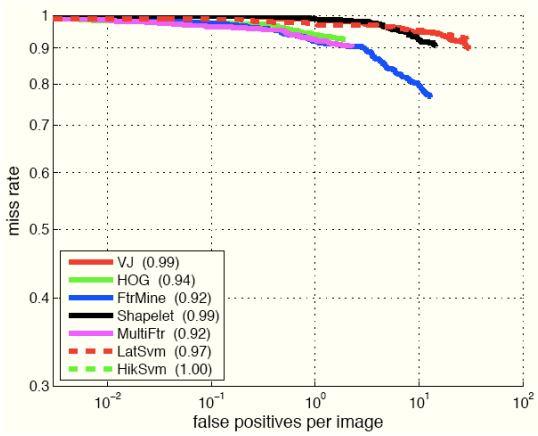
near scale



medium scale



far scale



[Dollar,Wojek,Perona,Schiele@CVPR-09]

Remaining Failure Cases (for INRIA-people dataset)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Missing Detections: [Wojek,Schiele@DAGM-08]



(a) Unusual articulation (b) Difficult contrast (c) Occlusion (d) Person carrying goods

- 149 missing detections:
  - 44 difficult contrast & backgrounds
  - 43 occlusion & carried bags
  - 37 unusual articulations
  - 18 over- / underexposure
  - 7 wrong scale (too small/large)

Remaining Failure Cases (for INRIA-people dataset)

TECHNISCHE UNIVERSITÄT DARMSTADT

- False Positives

[Wojek,Schiele@DAGM-08]



(e) Detection on parts (f) Too large scale (g) Detection on vertical structures (h) Cluttered background (i) Missing annotation

- 149 false positives:
  - 54 vertical structures / street signs
  - 31 cluttered background
  - 28 too small scale (body parts)
  - 24 too large scale detections
  - 12 people that are not annotated :-)

Motion HOG: Processing Chain

TECHNISCHE UNIVERSITÄT DARMSTADT

Detection windows

↓

Normalise gamma & colour

↓

Compute optical flow

↓

Compute differential flow

↓

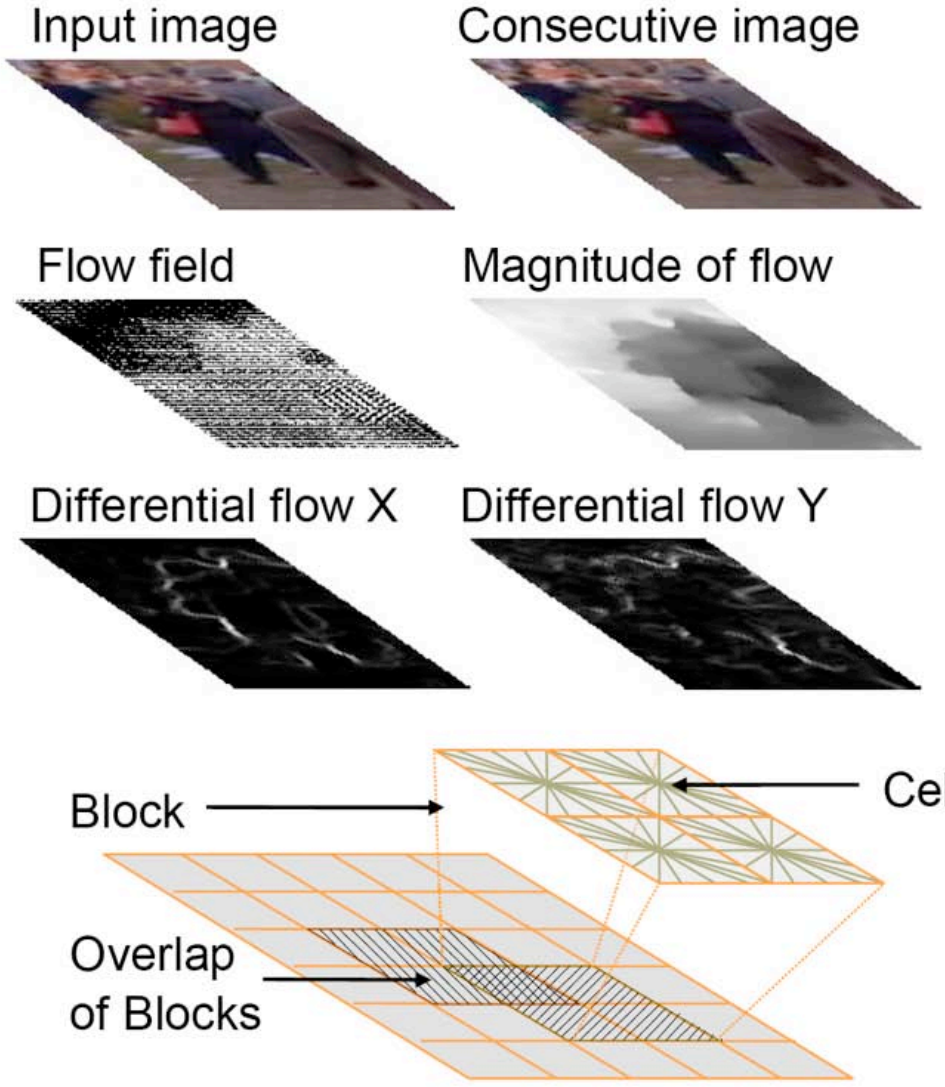
Accumulate votes for differential flow orientation over spatial cells

↓

Normalise contrast within overlapping blocks of cells

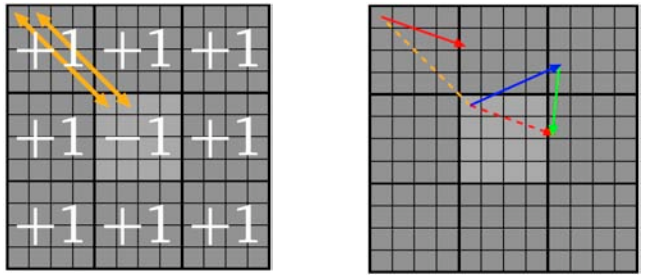
↓

Collect HOGs for all blocks over detection window



Motion HOG

- Encoding of the Optic Flow Differences:
  - here: Internal Motion Histogram (IMH)

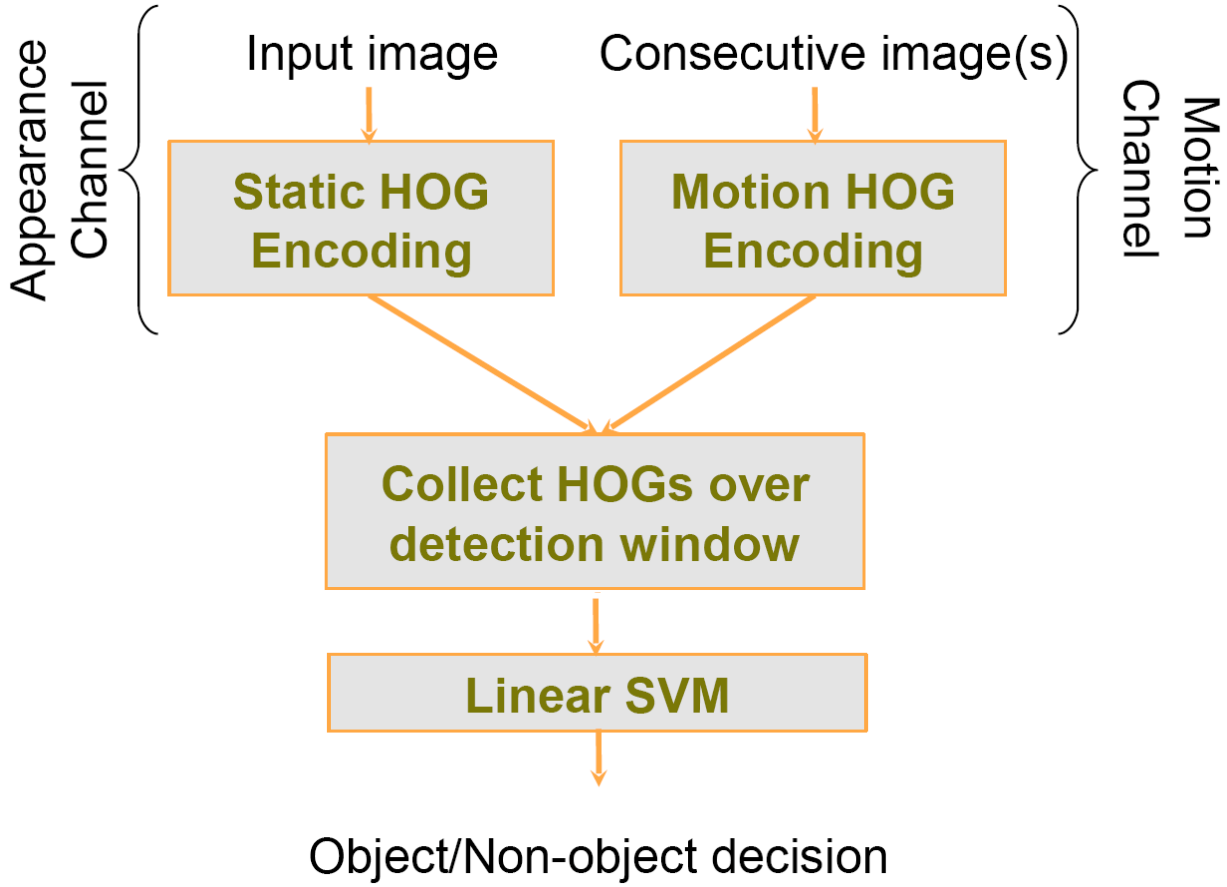


  - other possibility: Motion Boundary Histogram (MBH)
    - separte optical flow into its horizontal and vertical components



Combined HOG: Joint Modeling of Appearance & Motion

TECHNISCHE UNIVERSITÄT DARMSTADT

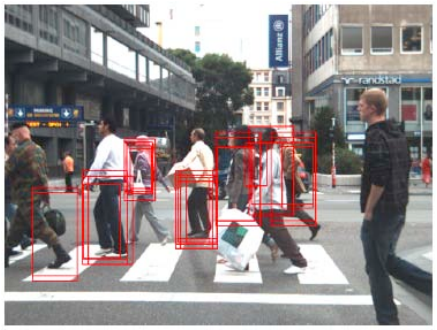


TUD-Brussels - Urban Onboard Dataset

TECHNISCHE UNIVERSITÄT DARMSTADT

- Training Set (image resolution 720x576 pixels)
  - positive set: 1092 image pairs (Darmstadt) (with 1776 pedestrians)
  - negative set: 192 image pairs
    - 85 pairs - taken from inner city (Darmstadt)
    - 107 pair recorded from moving car (Brussels)
    - to find hard examples: 26 image pairs with 183 pedestrians



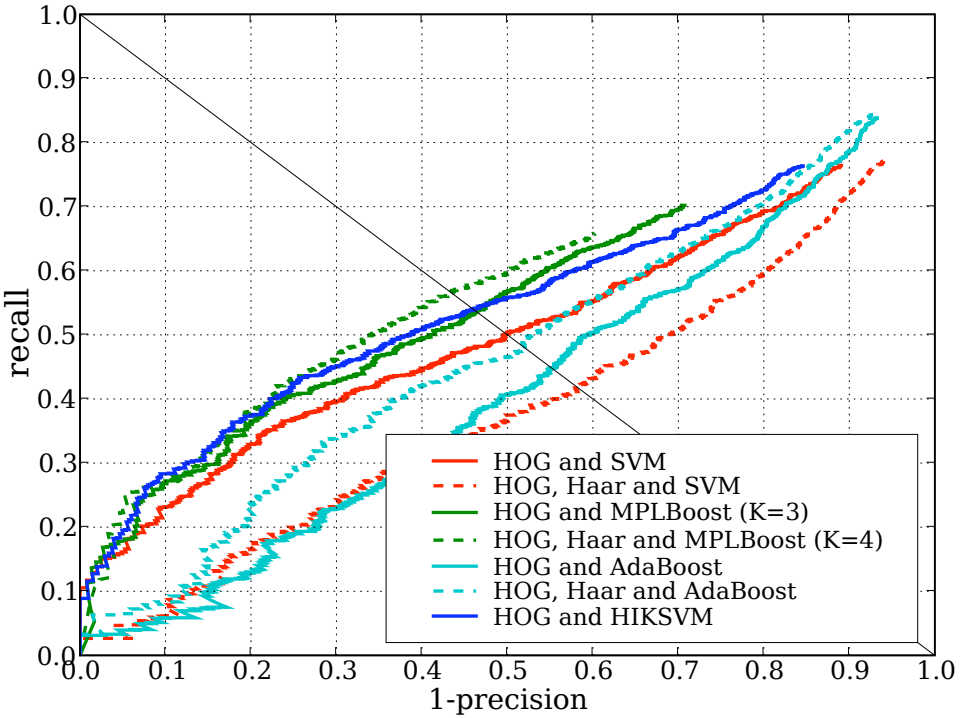


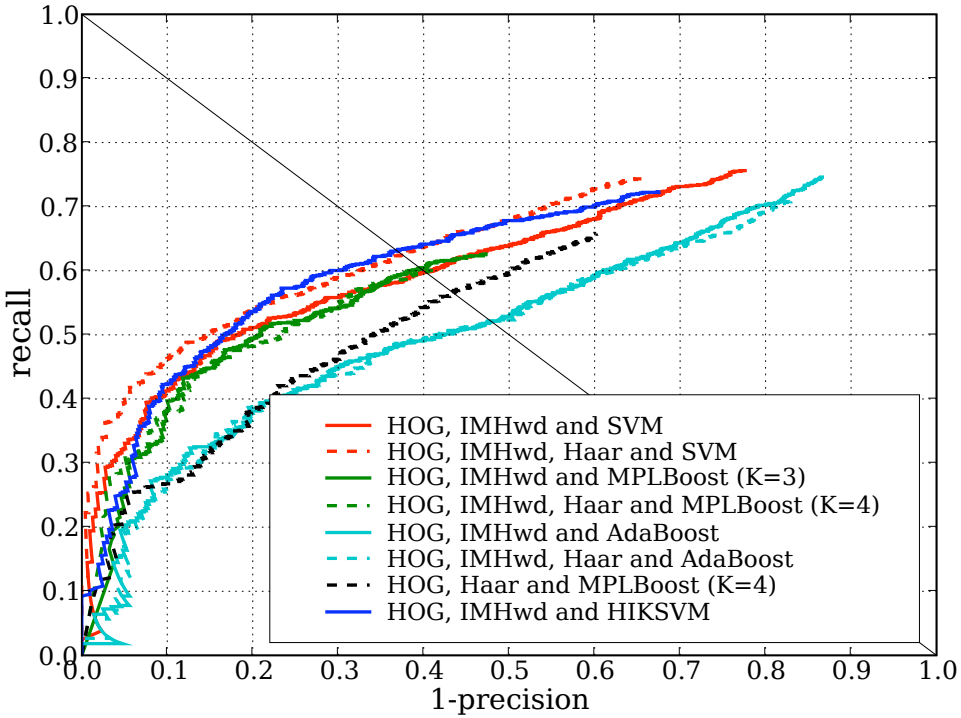
- Test Set (image resolution 640x480 pixels)
  - 508 image pairs with 1326 pedestrians (Brussels)

Quantitative Results
Static & Motion Features

TECHNISCHE UNIVERSITÄT DARMSTADT

- Most important Results:
  - motion features (right) outperform static features (left)
  - overall best: multi-cue + linear/Hik-SVM best
  - MPLBoost competitive to HikSVM for static features





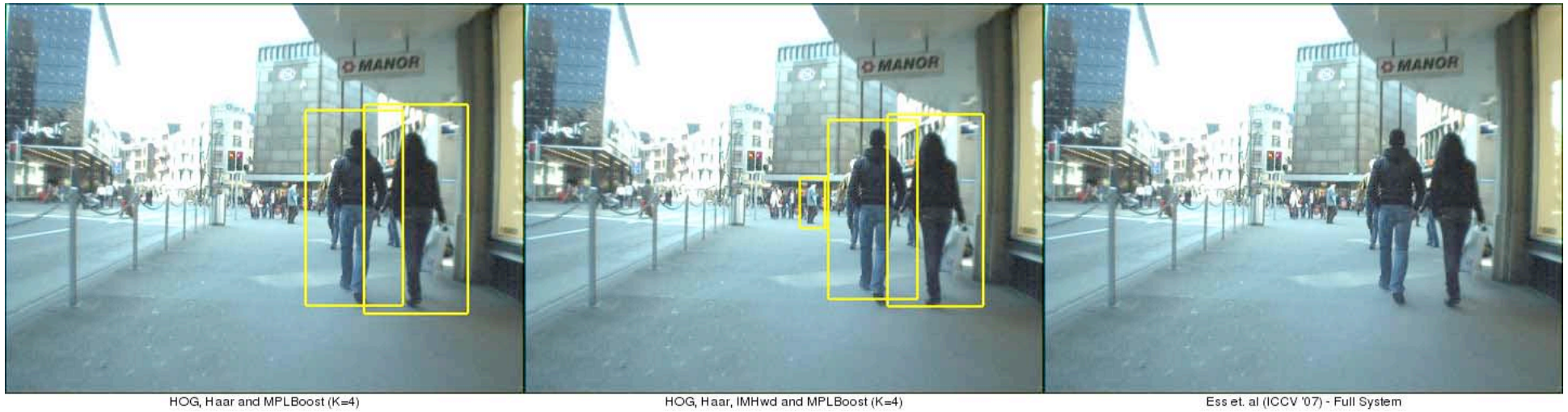
Motion-Based People Detection vs. full ETH-system

TECHNISCHE UNIVERSITÄT DARMSTADT

[Wojek,Walk,Schiele@CVPR-09]

- Comparison
  - our best static person detector
  - our best motion-based person detector
  - full ETH-Zurich system (Ess, Leibe, van Gool) using stereo, ground-plane, structure-from-motion, tracking, ...

our best static | our best motion | ETH full system



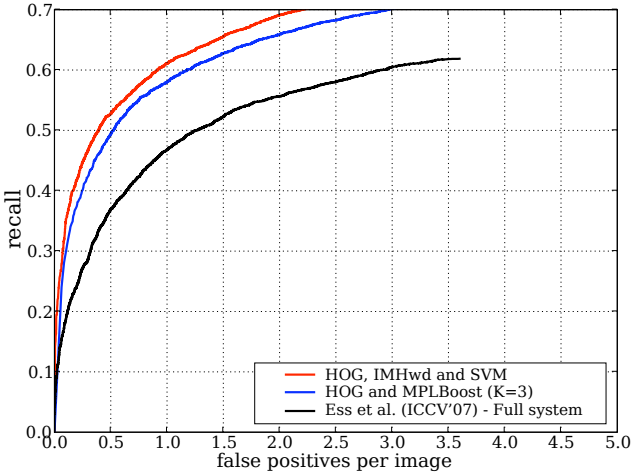
HOG, Haar and MPLBoost (K=4) | HOG, Haar, IMHwd and MPLBoost (K=4) | Ess et. al (ICCV '07) - Full System

Motion-Based People Detection vs. full ETH-system

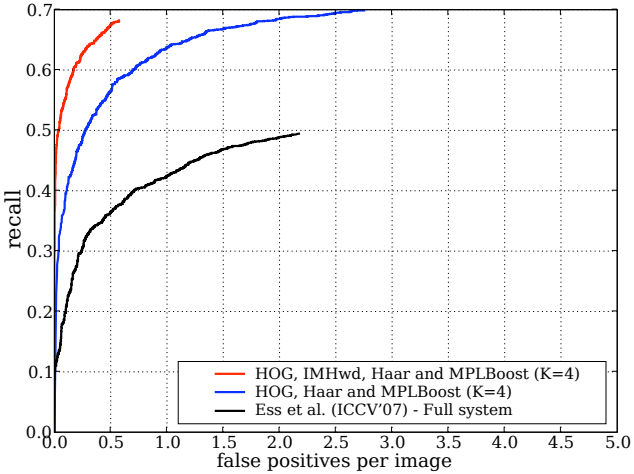
TECHNISCHE UNIVERSITÄT DARMSTADT

- Quantiative Comparison [Wojek,Walk,Schiele@CVPR-09]
  - blue line: our best static person detector
  - red line: our best motion-based person detector
  - black line: full ETH-Zurich system (Ess, Leibe, van Gool) using stereo, ground-plane, structure-from-motion, tracking, ...

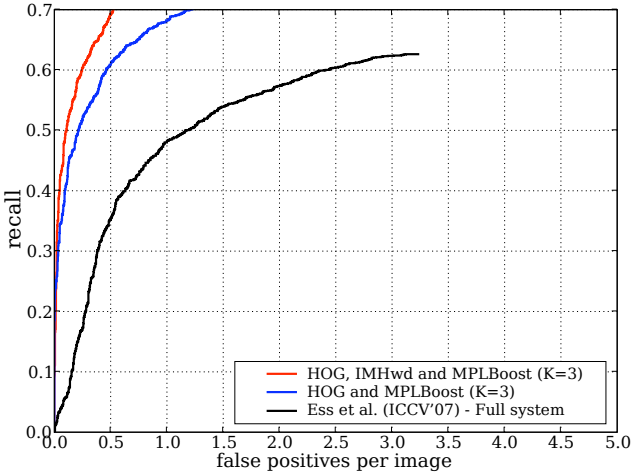
ETH-1



ETH-2



ETH-3

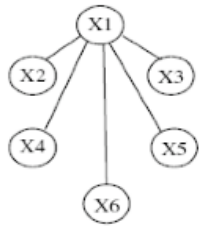


Part-Based Object and People Detection
Cognitive Science Summerschool 2oo9

TECHNISCHE UNIVERSITÄT DARMSTADT

Part 4: Part Based Models

e.g. ISM: flexible spatial relationships

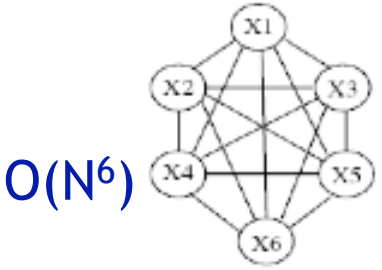


Bernt Schiele

Department of Computer Science
TU Darmstadt, Germany



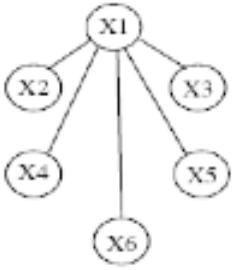
Different Connectivity Structures



$O(N^6)$

a) Constellation

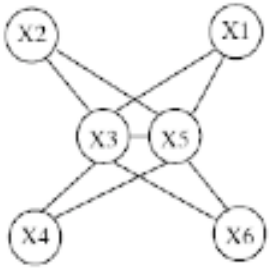
Fergus et al. ’03
Fei-Fei et al. ‘03



$O(N^2)$

b) Star shape

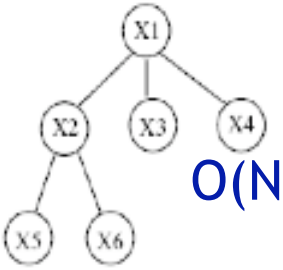
Leibe et al. ’04
Crandall et al. ‘05
Fergus et al. ’05



$O(N^3)$

c) k-fan ($k = 2$)

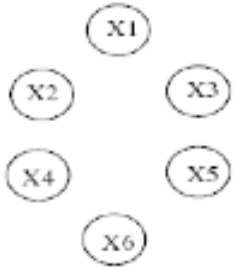
Crandall et al. ‘05



$O(N^2)$

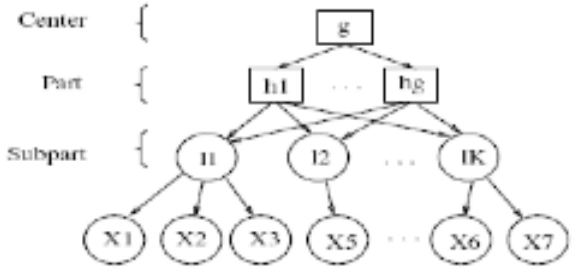
d) Tree

Felzenszwalb & Huttenlocher ‘00



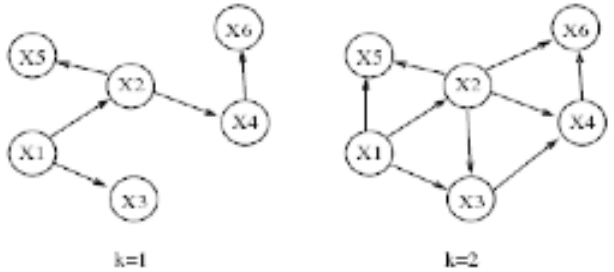
e) Bag of features

Csurka ’04
Vasconcelos ‘00



f) Hierarchy

Bouchard & Triggs ‘05



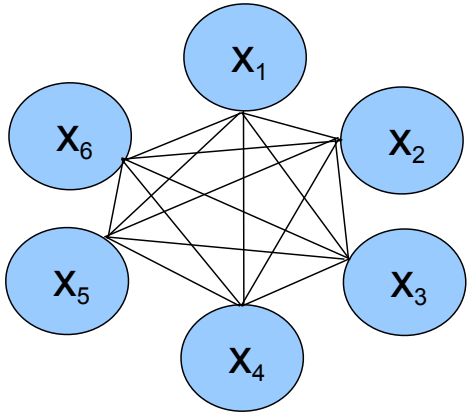
g) Sparse flexible model

Carneiro & Lowe ‘06

from [Carneiro & Lowe, ECCV’06]

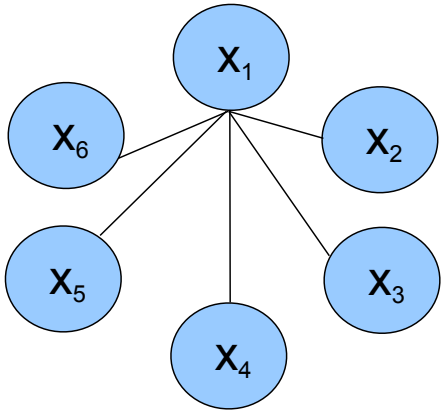
Spatial Models Considered Here (back in 2oo7)

Fully connected shape model



- e.g. Constellation Model
- Parts fully connected
- Recognition complexity: $O(N^P)$
- Method: Exhaustive search

"Star" shape model



- e.g. ISM
- Parts mutually independent
- Recognition complexity: O(NP)
- Method: Gen. Hough Transform

Slide credit: Rob Fergus

Overview of lecture parts 3 & 4...

TECHNISCHE UNIVERSITÄT DARMSTADT

- Global Feature Based Methods for People Detection (part 3)
  - A Performance Evaluation of Single and Multi-Feature People Detection [Wojek,Schiele@DAGM-08]
  - Pedestrian Detection: A New Benchmark [Dollar,Wojek,Perona,Schiele@CVPR-09]
  - Multi-Cue Onboard Pedestrian Detection [Wojek,Walk,Schiele@CVPR-09]
- Part-Based Model for People & Object Detection (part 4)
  - Detection by Tracking and Tracking by Detection [Andriluka,Roth,Schiele@CVPR-08]
  - Pictorial Structures Revisited: People Detection and Articulated Pose Estimation [Andriluka,Roth,Schiele@CVPR-09]
  - A Shape-Based Object Class Model for Knowledge Transfer [Stark,Goesele,Schiele@ICCV-09]

Motivation: People Detection and Tracking

TECHNISCHE UNIVERSITÄT DARMSTADT

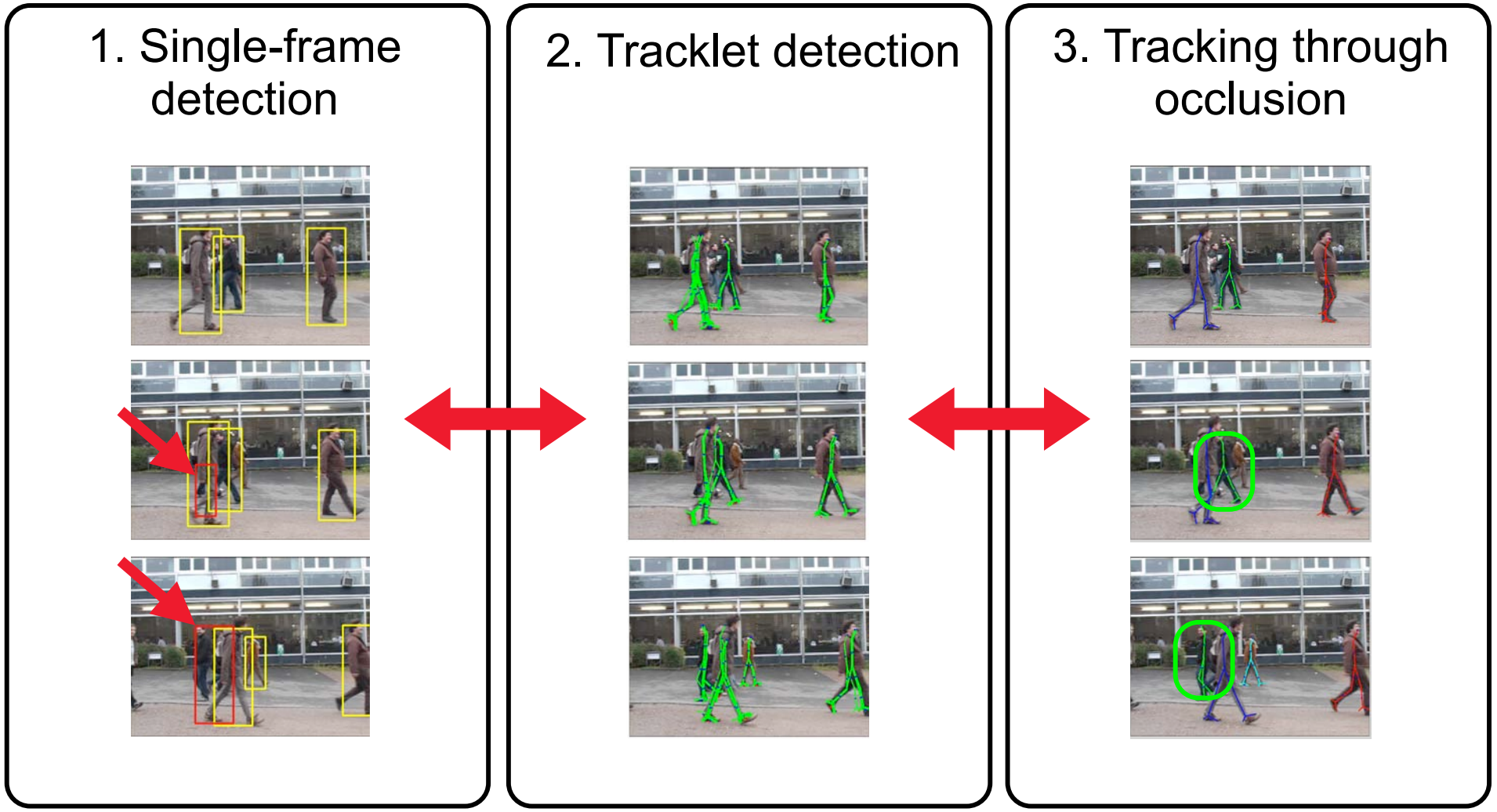
- Goal: Detection and tracking of people in complex scenes
- Challenges for Detection:
  - Partial occlusions
  - Appearance variation
  - Data association difficult
- Challenges for Tracking:
  - Dynamic backgrounds
  - Multiple people
  - Frequent long term occlusions



Overview

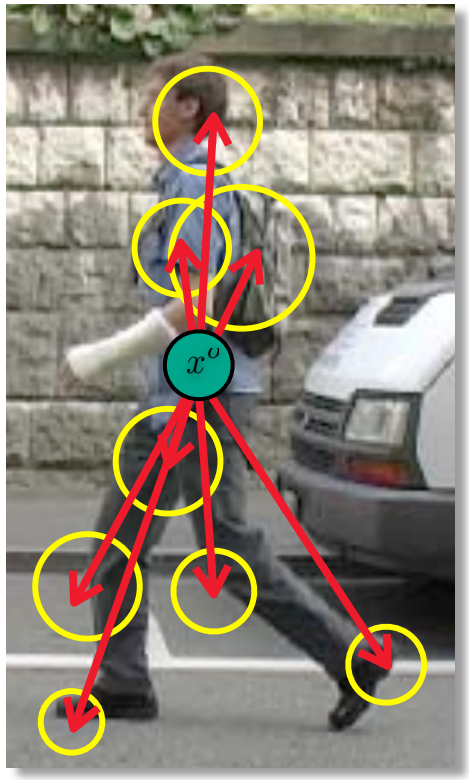
TECHNISCHE UNIVERSITÄT DARMSTADT

Three stages of our multi-person detection and tracking system:

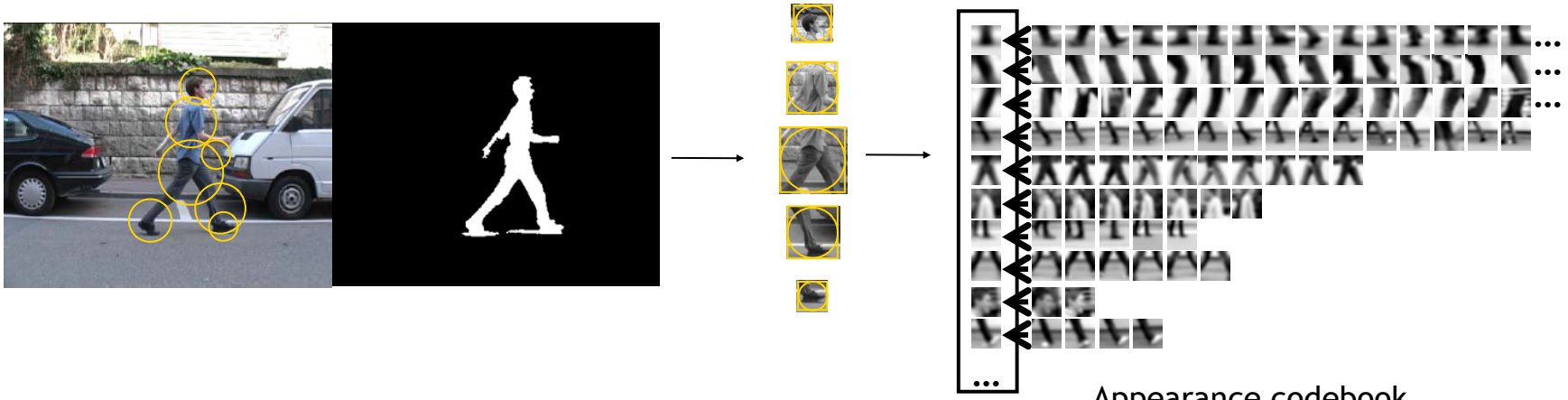


Single-frame Detector: partISM

- Appearance of parts:
  Implicit Shape Model (ISM)
  [Leibe, Seemann & Schiele, CVPR 2005]

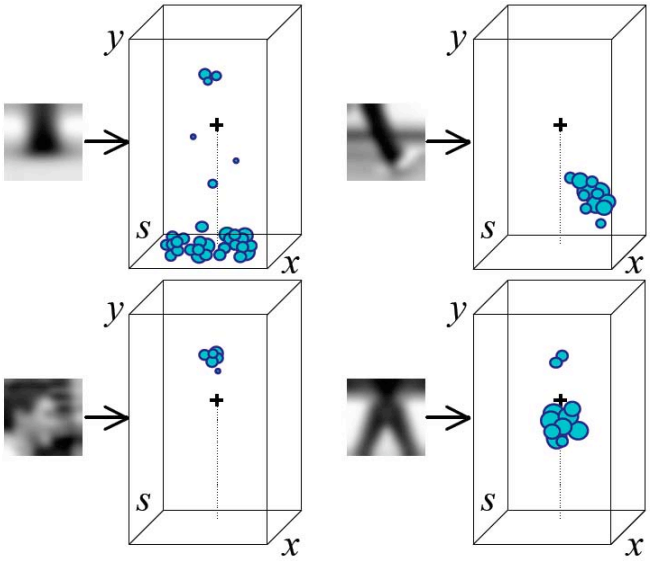


Implicit Shape Model - Representation



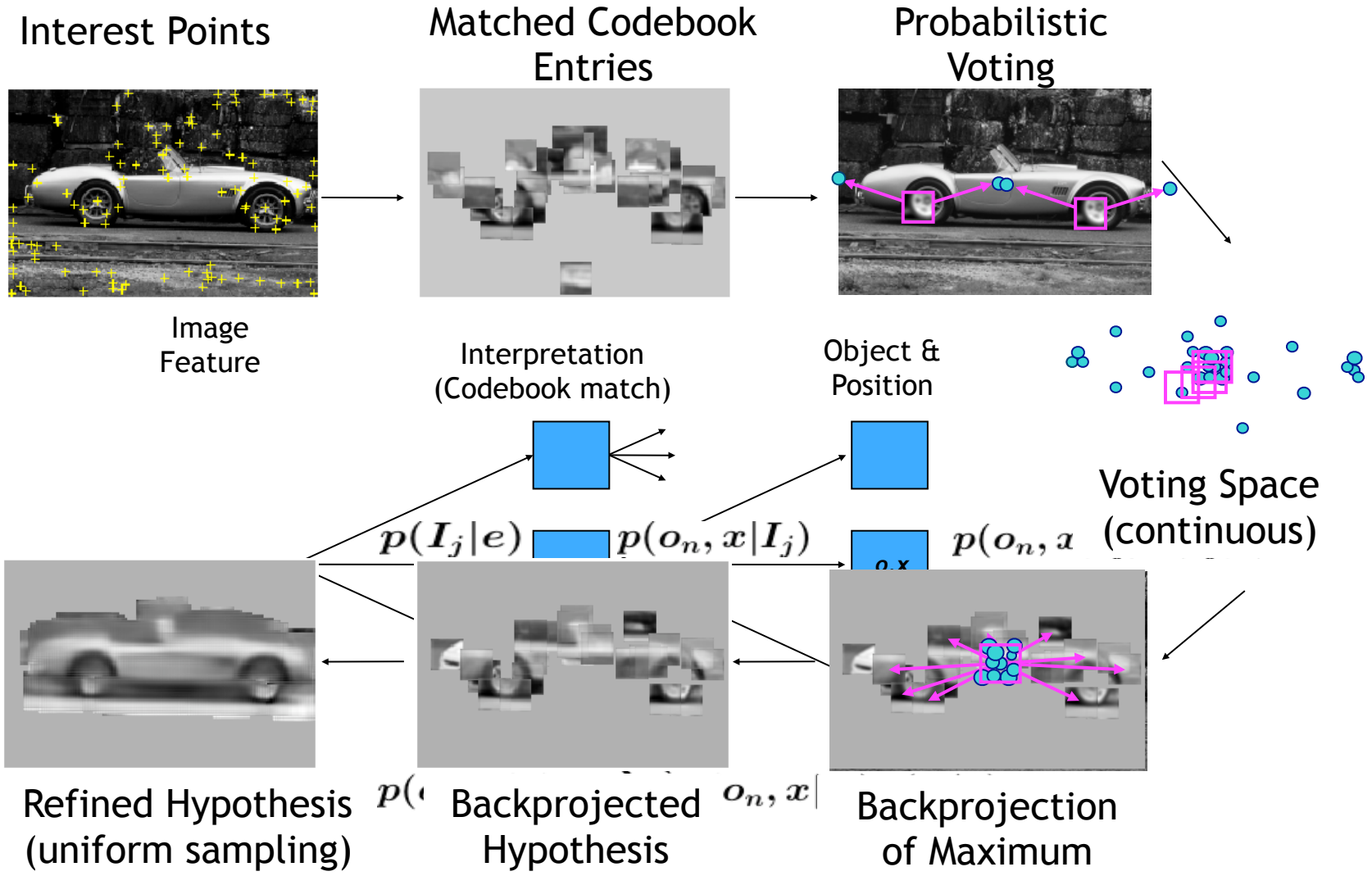
Appearance codebook

- Learn appearance codebook
  - Extract features at interest points (e.g. DoG)
  - Agglomerative clustering ⇒ codebook
- Learn codebook distributions (position & scale)
  - Match codebook to training images
  - Record matching positions on object

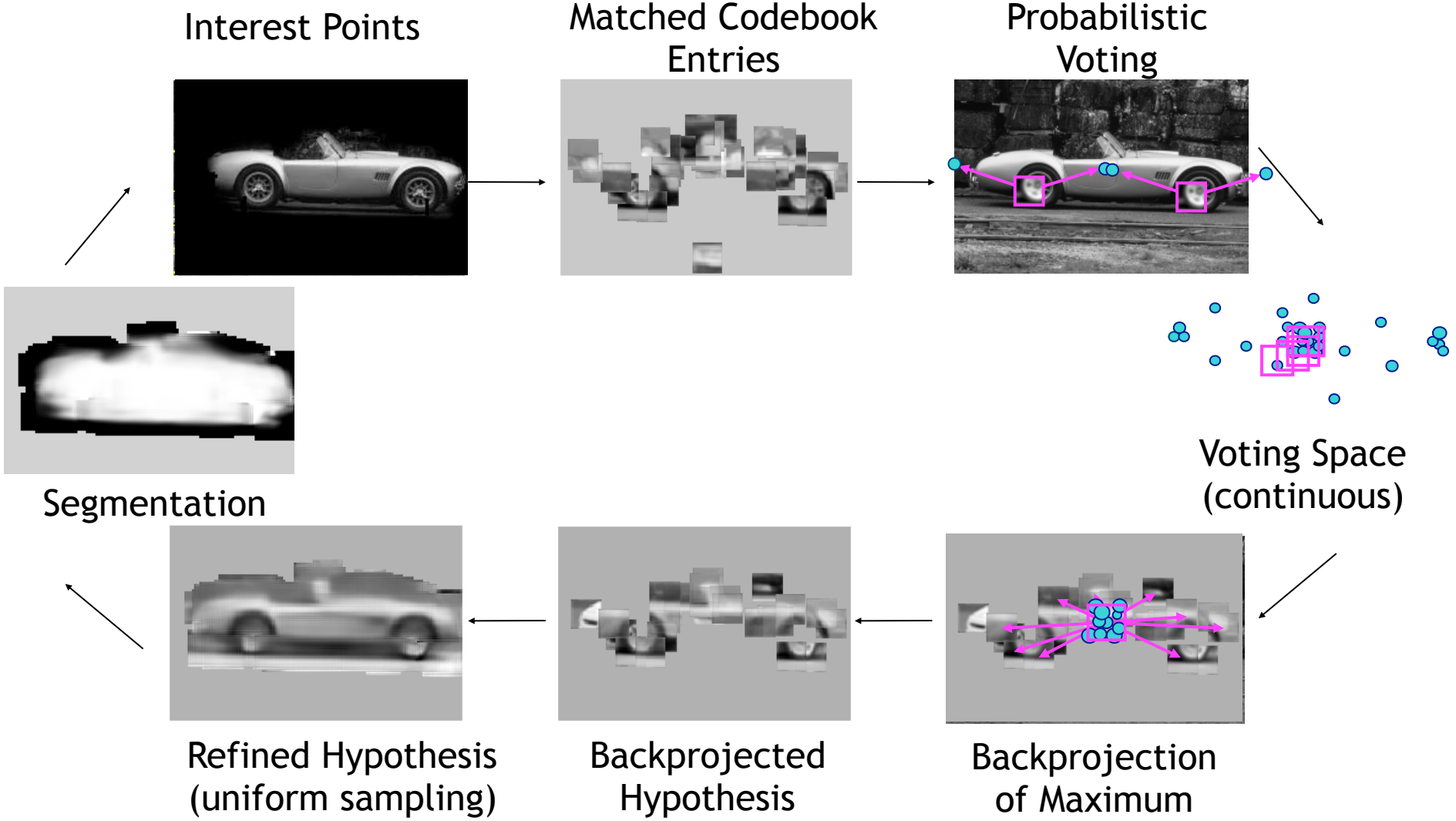


Spatial occurrence distributions

Object Class Recognition Procedure



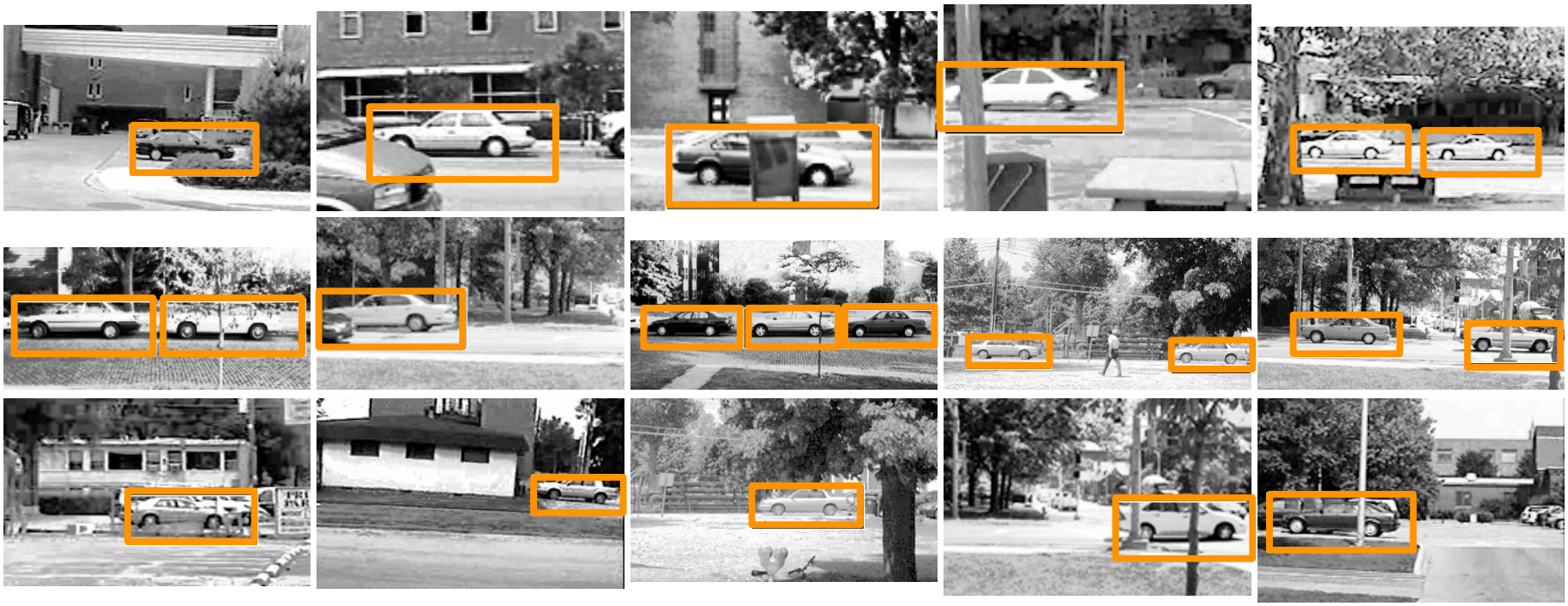
Categorization: “Closing the Loop”



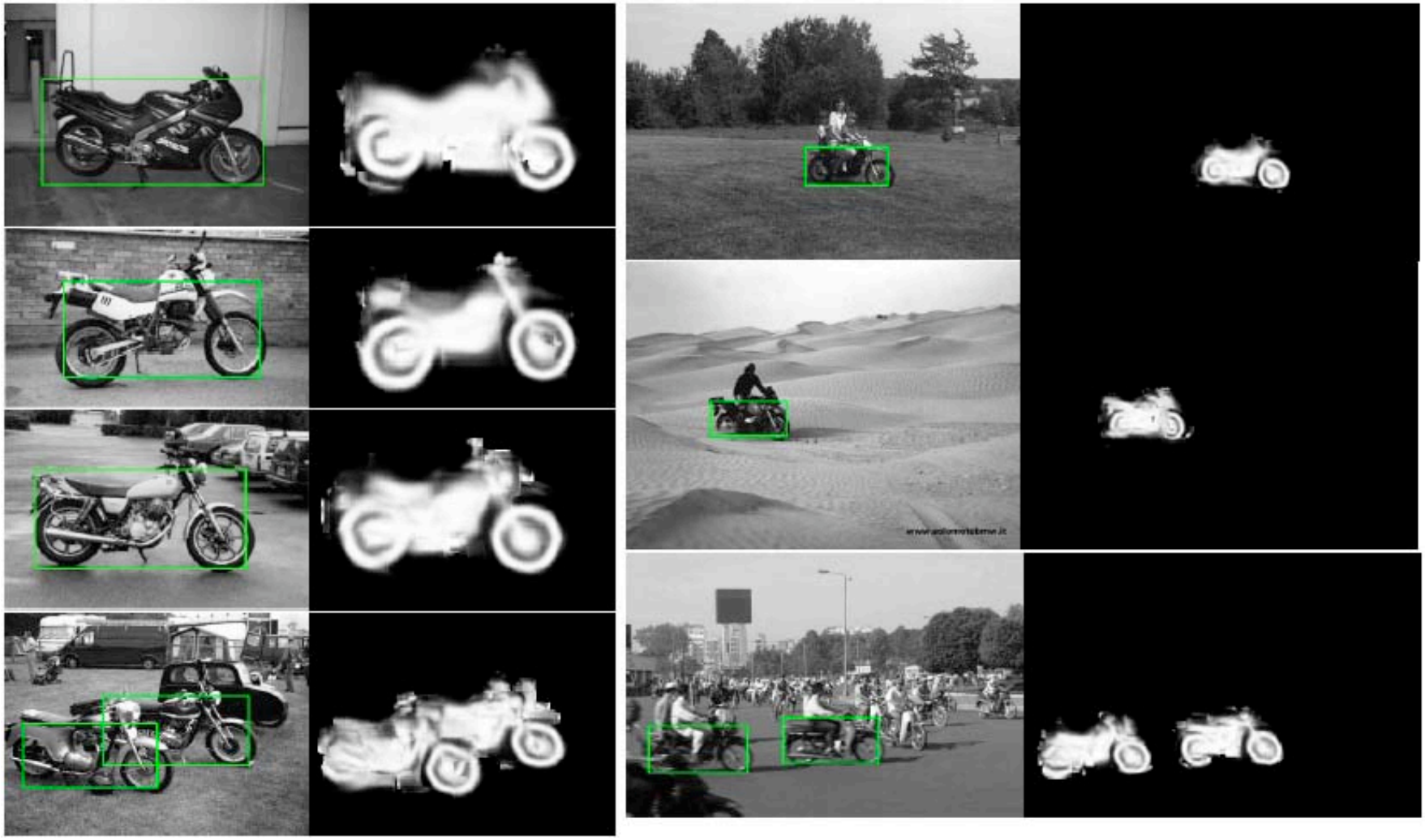
Detection Results

TECHNISCHE UNIVERSITÄT DARMSTADT

- Qualitative Performance (UIUC database - 200 cars)
  - Recognizes different kinds of cars
  - Robust to clutter, occlusion, low contrast, noise

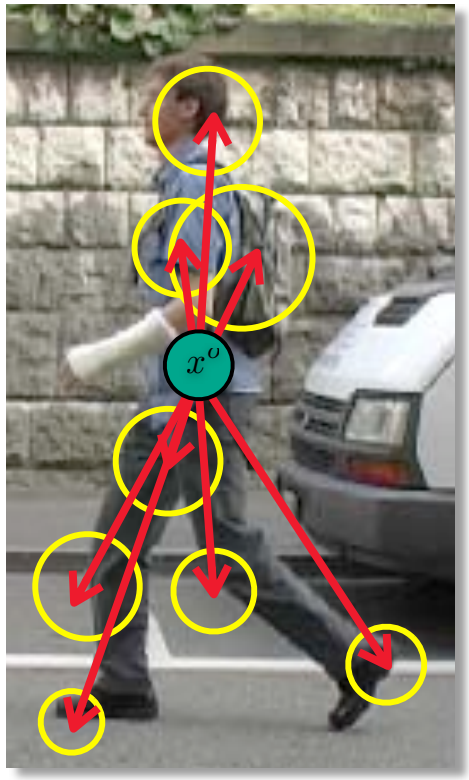


Example Category: Motorbikes



Single-frame Detector: partISM

- Appearance of parts:
  Implicit Shape Model (ISM)
  [Leibe, Seemann & Schiele, CVPR 2005]

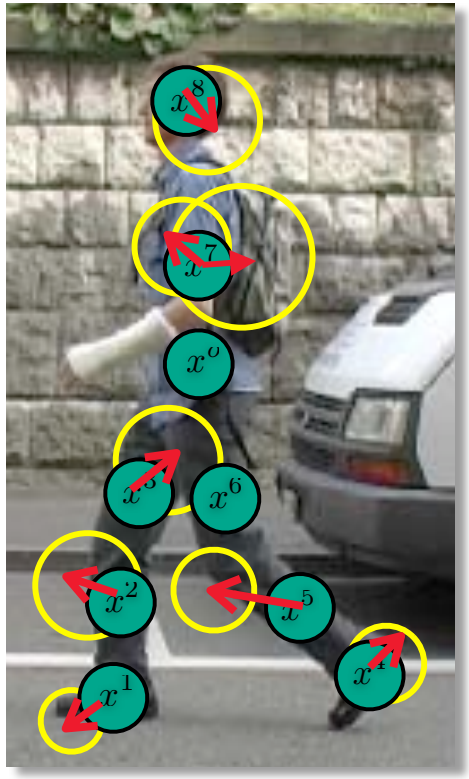


Single-frame Detector: partISM

- Appearance of parts:
  Implicit Shape Model (ISM)
  [Leibe, Seemann & Schiele, CVPR 2005]
- Part decomposition and inference:
  Pictorial structures model
  [Felzenszwalb & Huttenlocher, IJCV 2005]

$$p(L|E) \propto p(E|L)p(L)$$

Body-part positions Image evidence

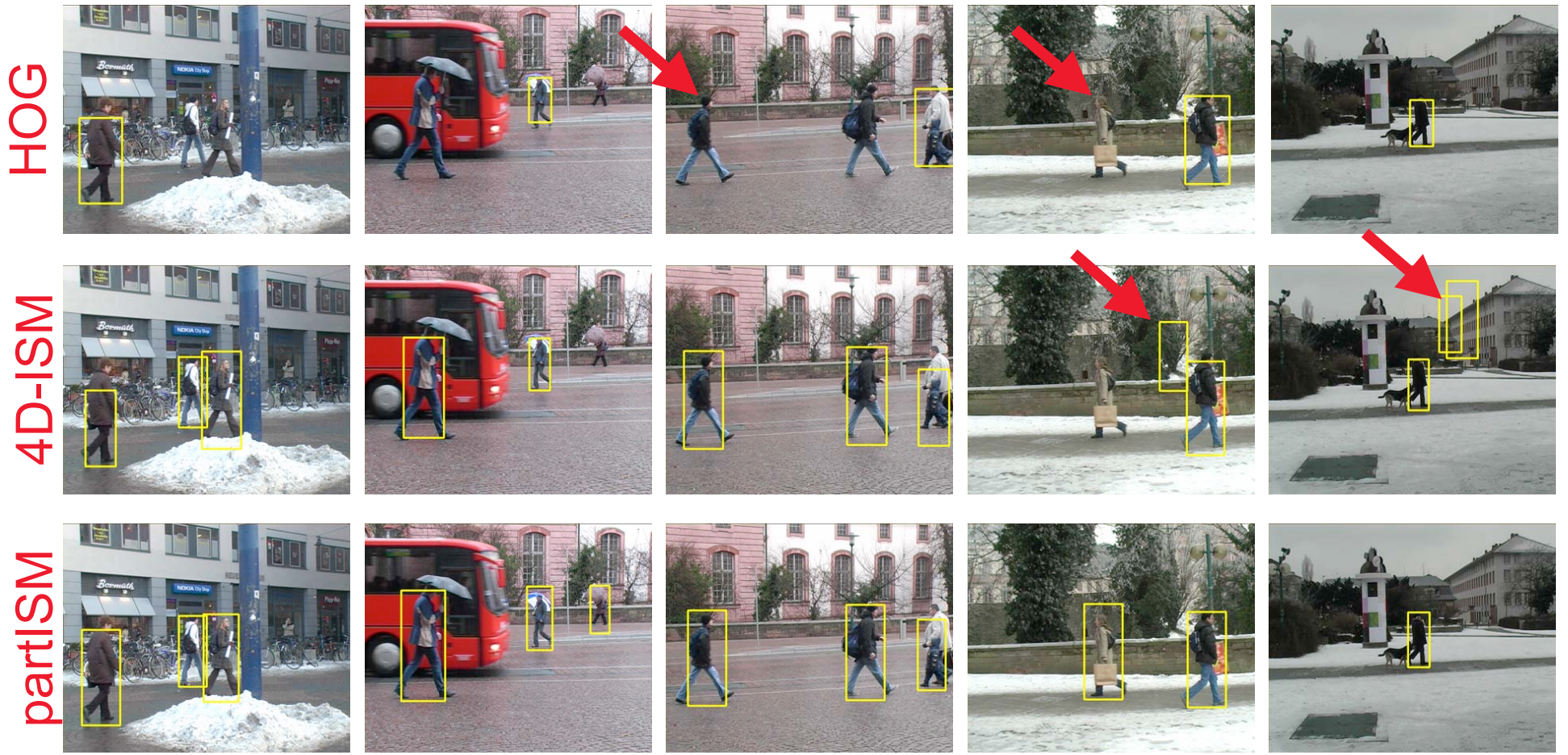


$$L = \{x^o, x^1, \ldots, x^8\}$$

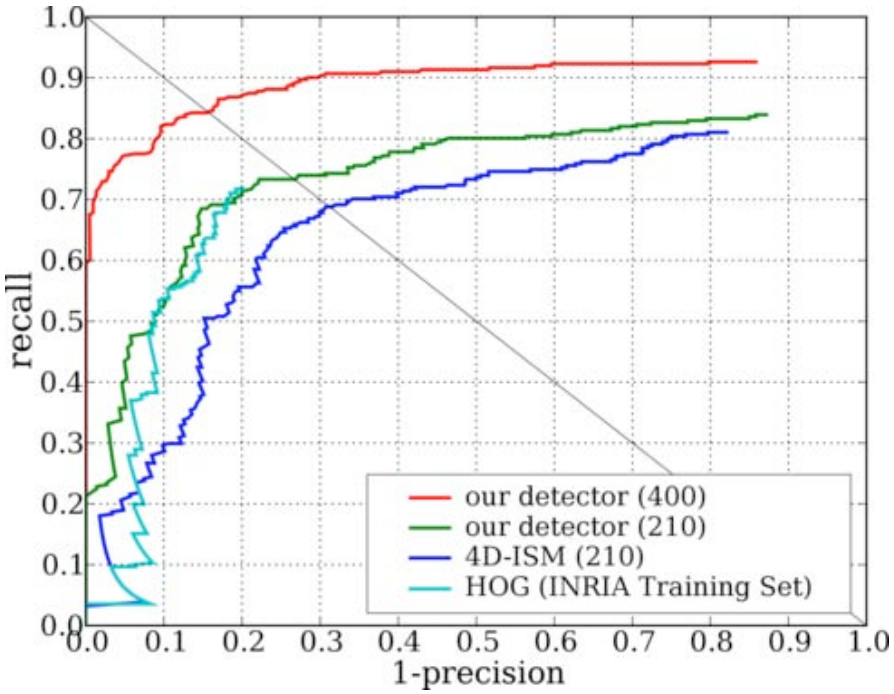
Single Frame Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- Detections at equal error rate:



Single-frame Detection Results



TUD pedestrians data

No occlusions

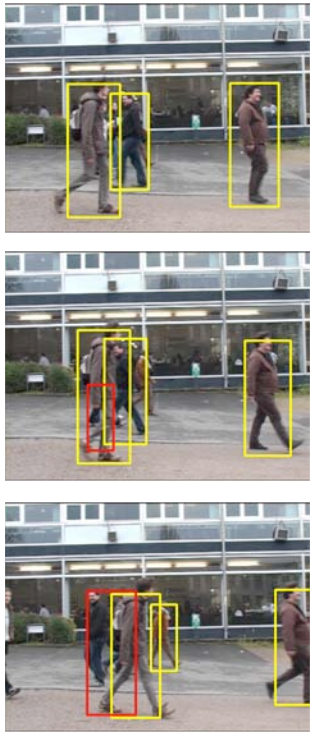
- partISM clearly outperforms 4D-ISM [Seemann et al, DAGM'06].
- Outperforms HOG [DalaI&Triggs, CVPR'05] with much less training data (Note: we only use sideviews).

Overview

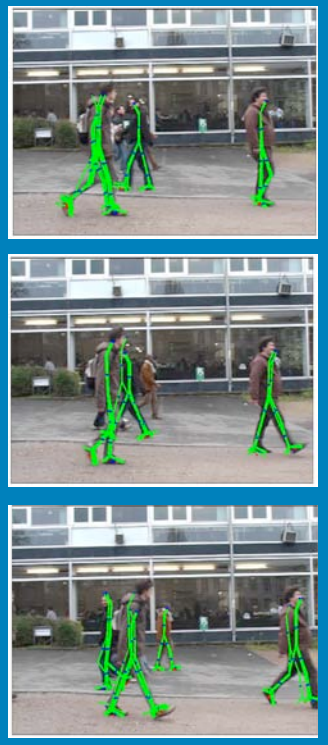
TECHNISCHE UNIVERSITÄT DARMSTADT

Three stages of our multi-person detection and tracking system:

1. Single-frame detection



2. Tracklet detection



3. Tracking through occlusion



Tracklet Detection in Short Subsequences

- Given: $E = [E_1, \ldots, E_m]$



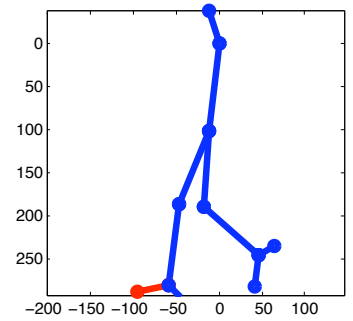
overlapping subsequences

- Want:



$\mathbf{X}^{o*} = [\mathbf{x}_1^{o*}, \ldots, \mathbf{x}_m^{o*}]$
body positions

$\mathbf{Y}^* = [\mathbf{y}_1^*, \ldots, \mathbf{y}_m^*]$
body configurations



- Posterior over positions and configurations:

$$p(\mathbf{X}^{o*}, \mathbf{Y}^* | E) \propto p(E | \mathbf{X}^{o*}, \mathbf{Y}^*) p(\mathbf{X}^{o*}) p(\mathbf{Y}^*).$$

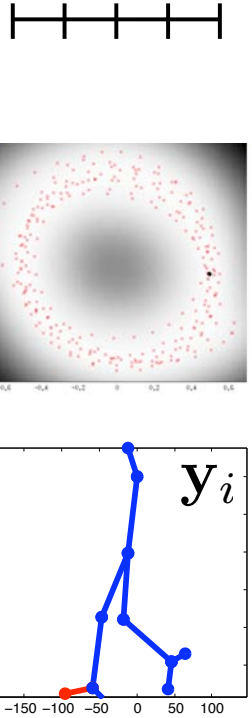
Likelihood model (partISM)

speed model (Gaussian) here: constant speed

dynamical body model (hGPLVM)

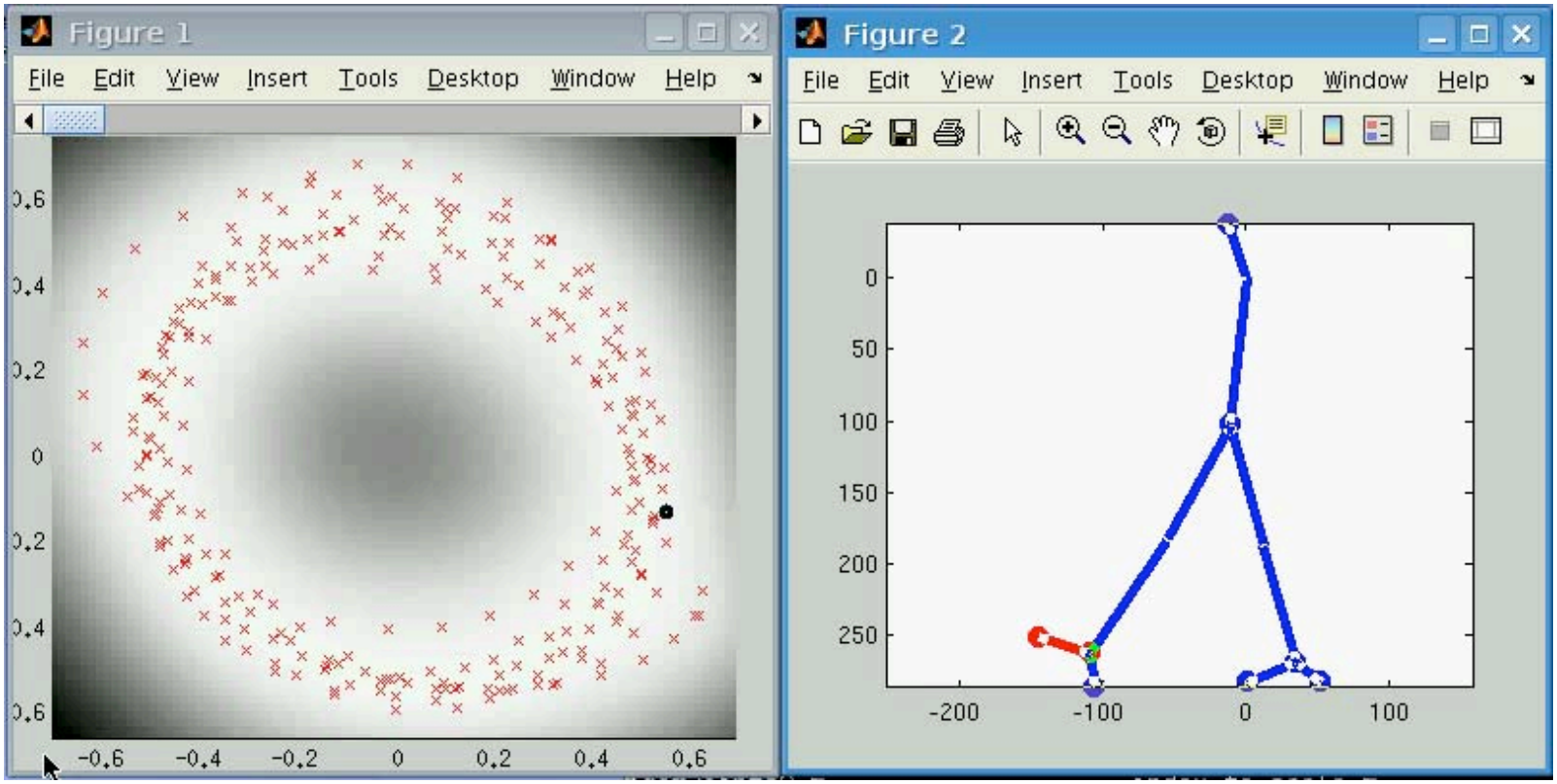
Modeling Body Dynamics

- $\mathbf{Y}^*$ is high-dimensional: Full body poses in m frames.
- Model the body dynamics using hierarchical Gaussian process latent variable model (hGPLVM) [Lawrence&Moore, ICML 2007]



Modeling Body Dynamics

- Visualization of the hierarchical Gaussian process latent variable model (hGPLVM)



Single-Frame Detector vs. Tracklet Detector

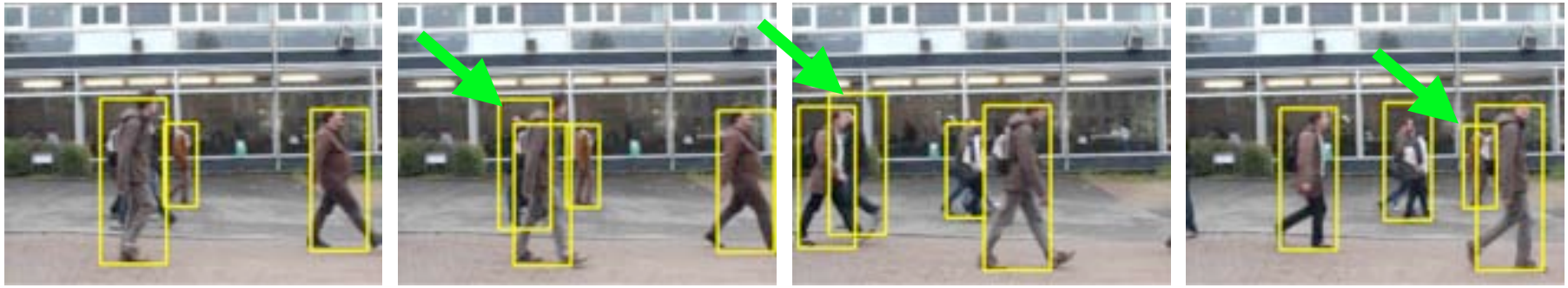
TECHNISCHE UNIVERSITÄT DARMSTADT

- At equal error rate:

partISM



Tracklet detector



- Fewer false positives.
- More robust detection of partially occluded people.

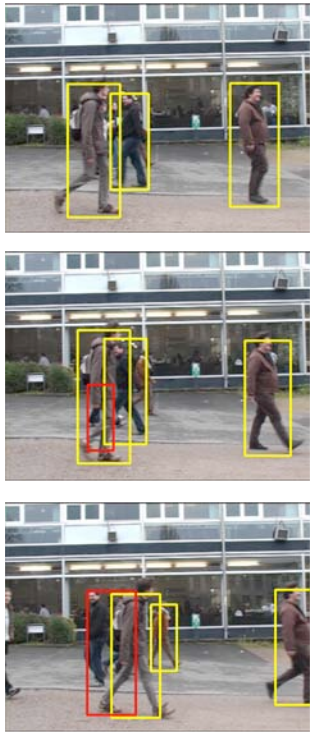
[Andriluka,Roth,Schiele@CVPR-08]

Overview

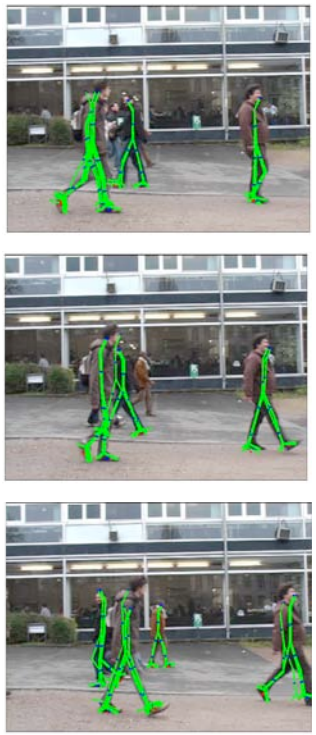
TECHNISCHE UNIVERSITÄT DARMSTADT

Three stages of our multi-person detection and tracking system:

1. Single-frame detection



2. Tracklet detection



3. Tracking through occlusion



Occlusion Recovery

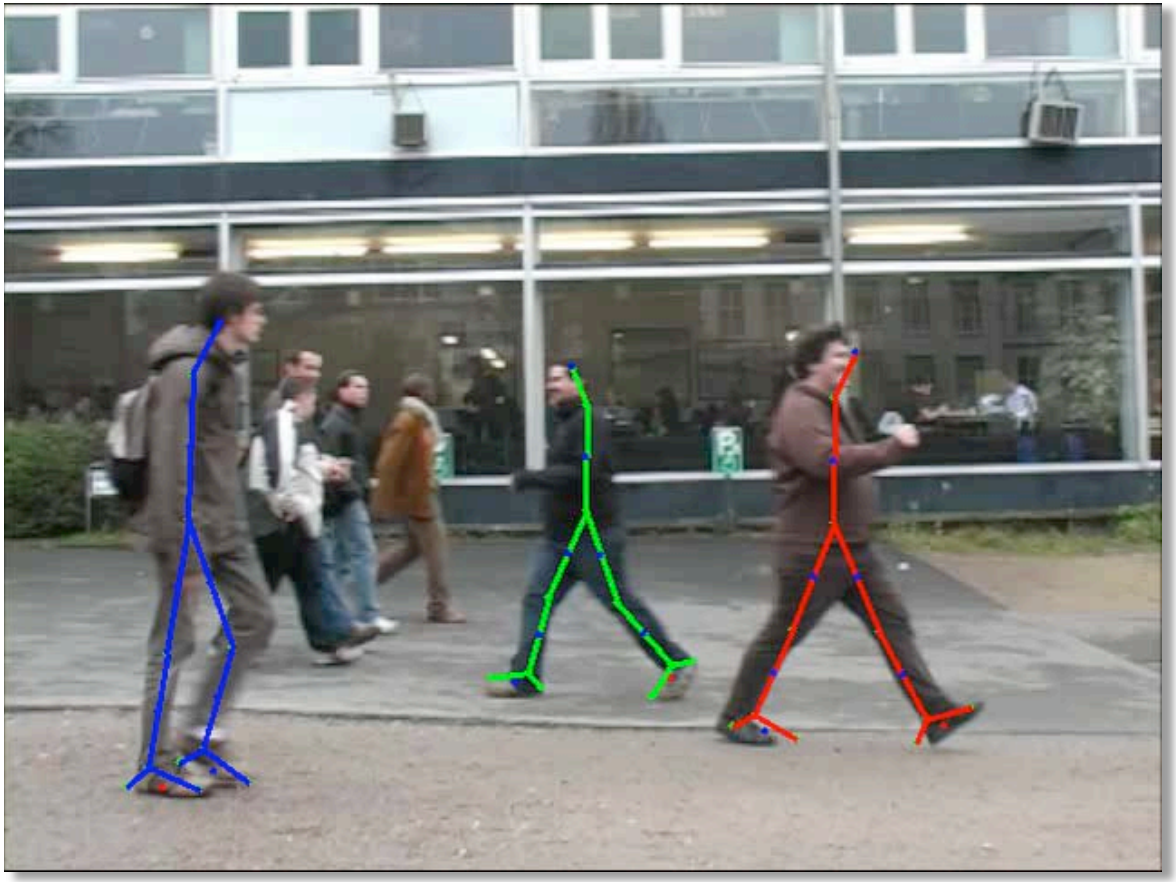


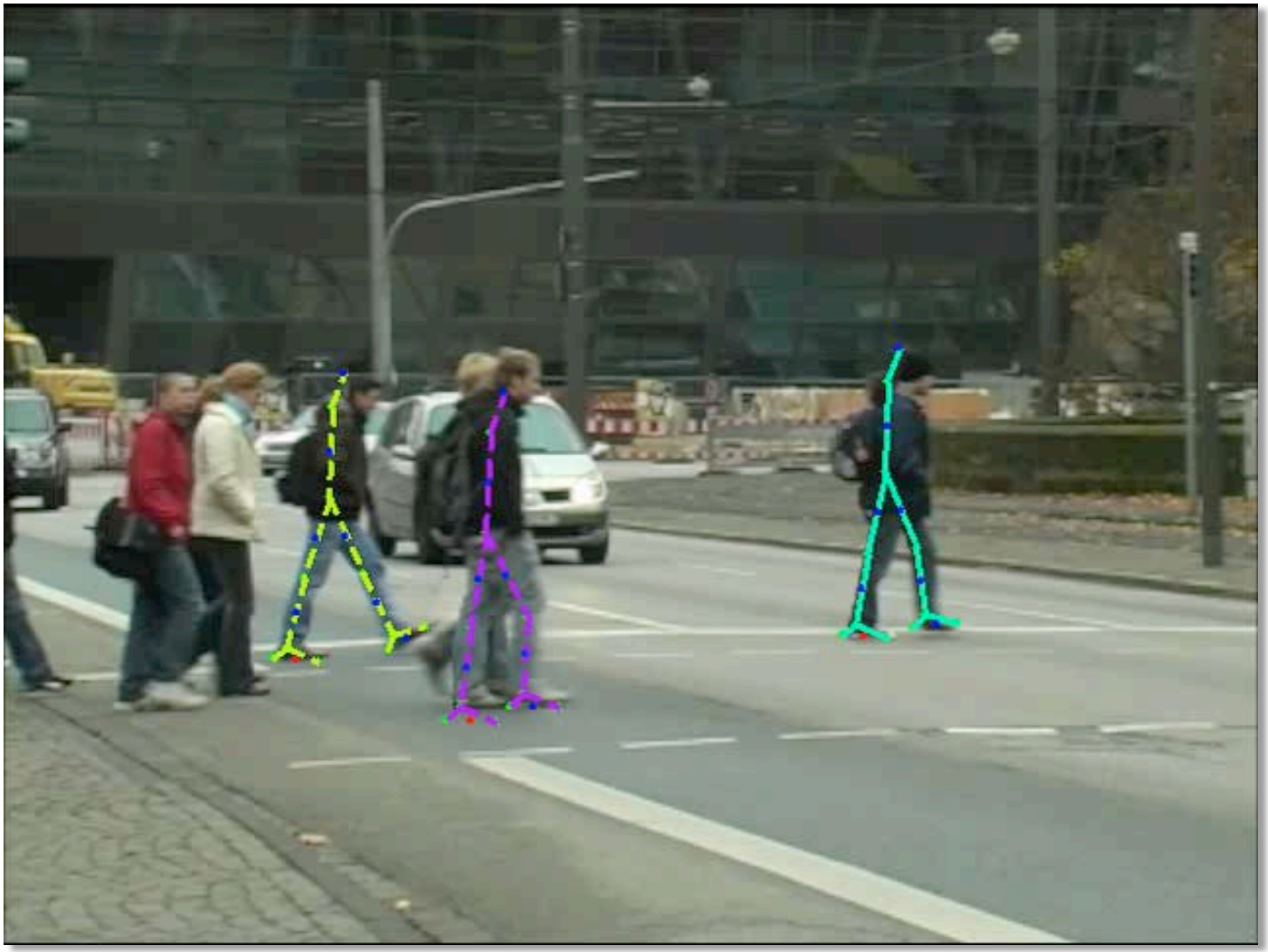
time ⟶

- Greedily link partial tracks based on:
  - Motion & articulation compatibility.
  - Plus appearance compatibility.

Occlusion Recovery

- Greedily link partial tracks based on:
  - Motion & articulation compatibility.
  - Plus appearance compatibility.





Pictorial Structure Revisited: Model Components [cvpr09]

TECHNISCHE UNIVERSITÄT DARMSTADT

Appearance Model:



Local Features

orientation 1

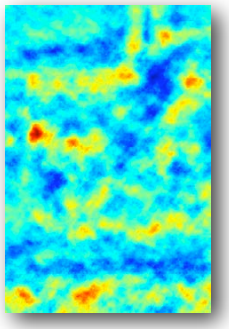


orientation K

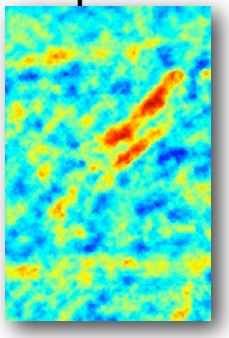


AdaBoost

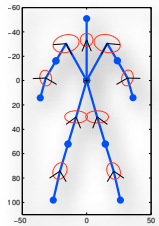
likelihood of part 1



likelihood of part N



Prior and Inference:



estimated pose



part posteriors



Model Components

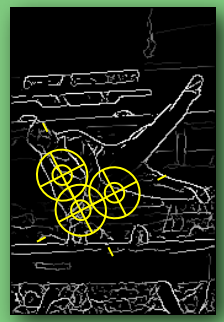
TECHNISCHE UNIVERSITÄT DARMSTADT

Appearance Model:

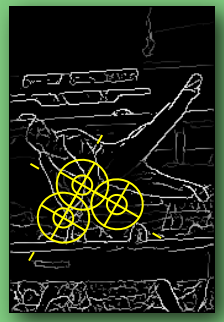


Local Features

orientation 1

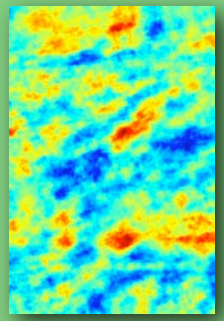


orientation K

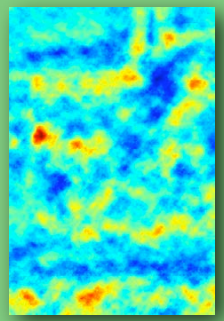


AdaBoost

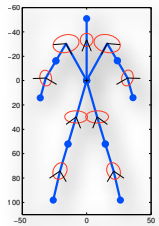
likelihood of part 1



likelihood of part N



Prior and Inference:



estimated pose

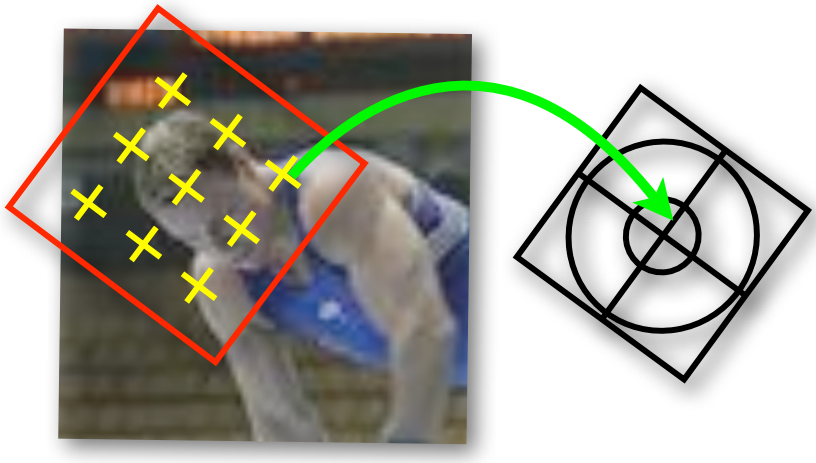


part posteriors



Likelihood Model

- Build on recent advances in object detection:
  - state-of-the-art image descriptor: Shape Context
    [Belongie et al., PAMI'02; Mikolajczyk&Schmid, PAMI'05]
  - dense representation
  - discriminative model: AdaBoost classifier for each body part



- Shape Context: 96 dimensions (4 angular, 3 radial, 8 gradient orientations)

- Feature Vector: concatenate the descriptors inside part bounding box
- head: 4032 dimensions
- torso: 8448 dimensions

Likelihood Model

- Part likelihood derived from the boosting score:

decision stump weight

decision stump output

$$\tilde{p}(\mathbf{d}_i|\mathbf{l}_i) = \max\left(\frac{\sum_t \alpha_{i,t} h_t(\mathbf{x}(\mathbf{l}_i))}{\sum_t \alpha_{i,t}}, \varepsilon_0\right)$$

part location

small constant to deal with part occlusions

Likelihood Model

TECHNISCHE UNIVERSITÄT DARMSTADT

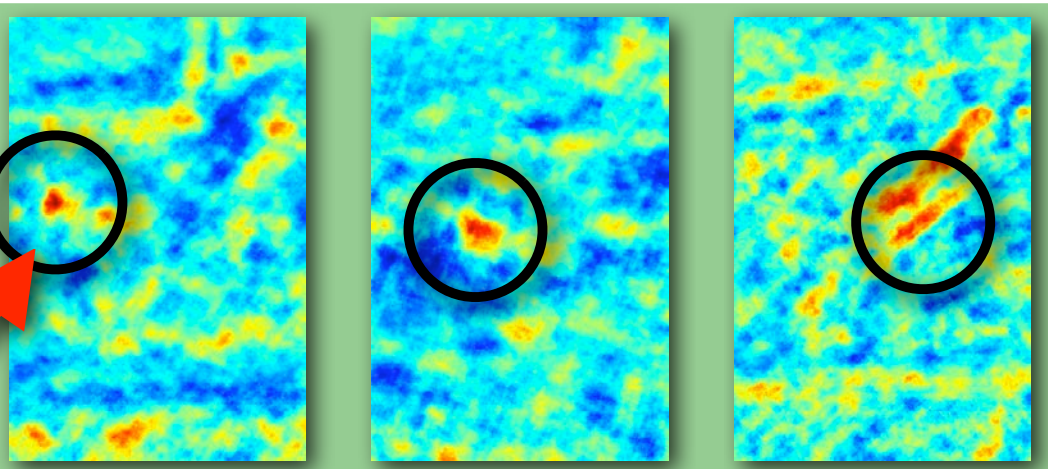
Input image



Head Torso Upper leg

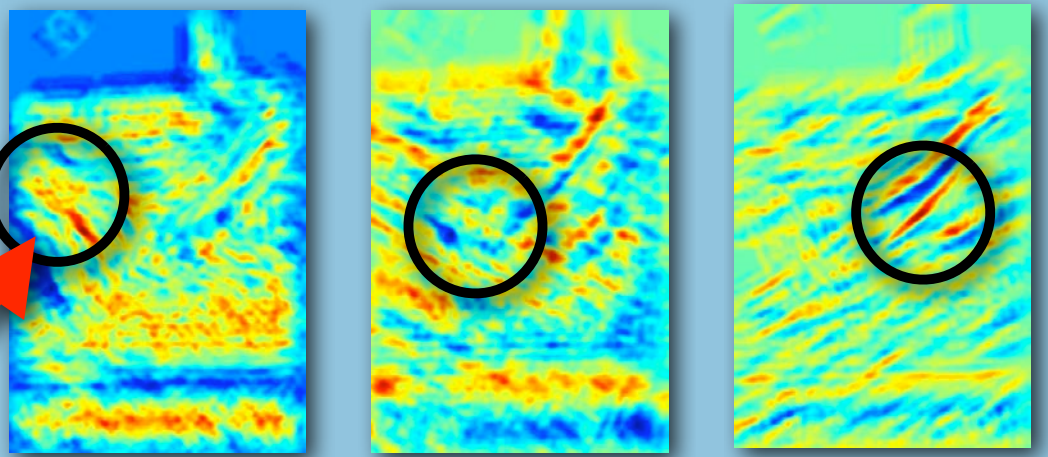
Our part likelihoods





[Ramanan, NIPS'06]





Likelihood Model

TECHNISCHE UNIVERSITÄT DARMSTADT

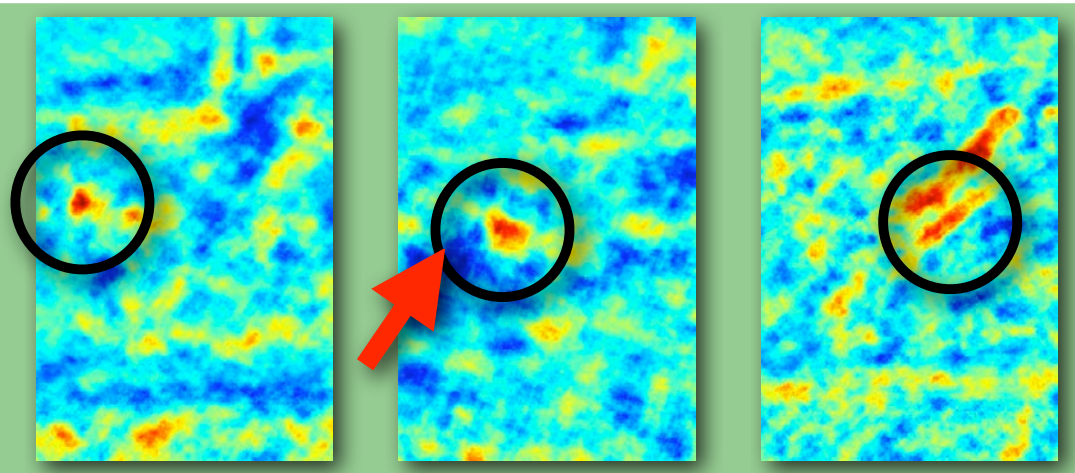
Input image



Head Torso Upper leg

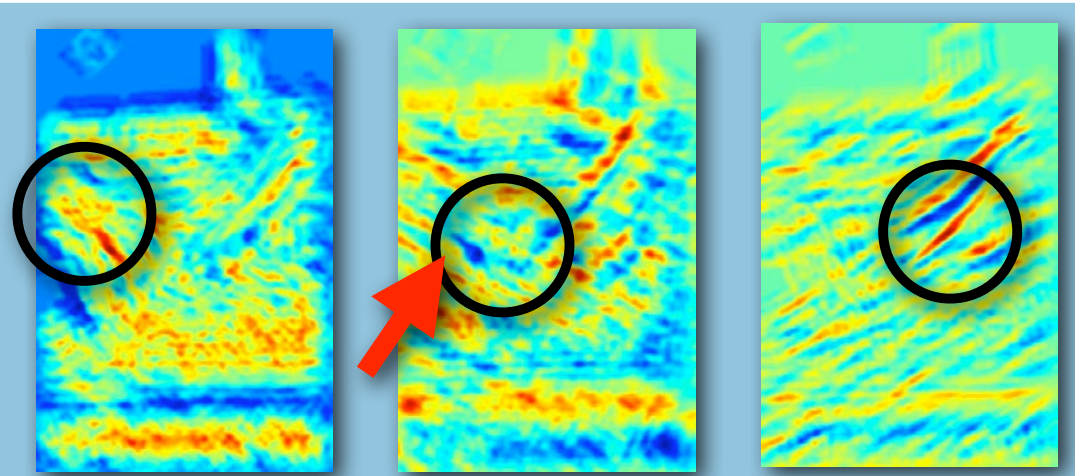
Our part likelihoods





[Ramanan, NIPS'06]





Likelihood Model

TECHNISCHE UNIVERSITÄT DARMSTADT

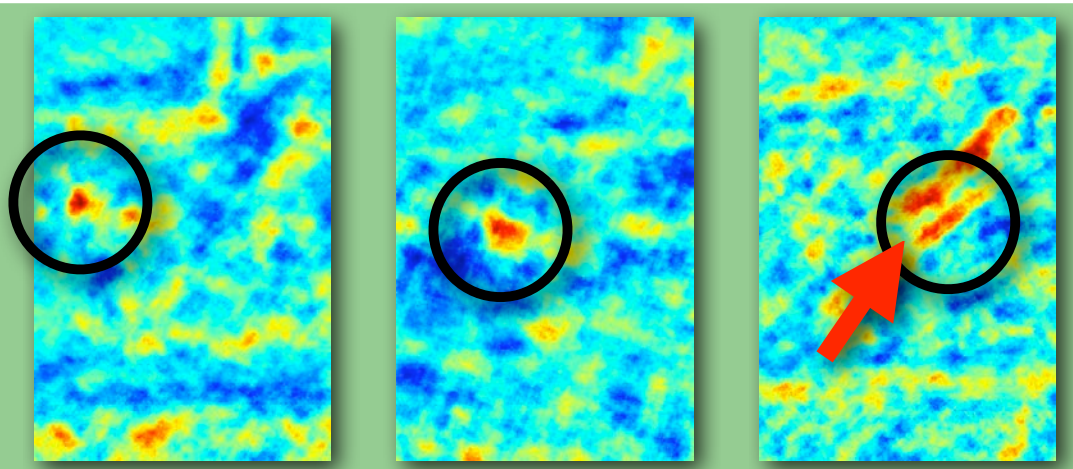
Input image



Head Torso Upper leg

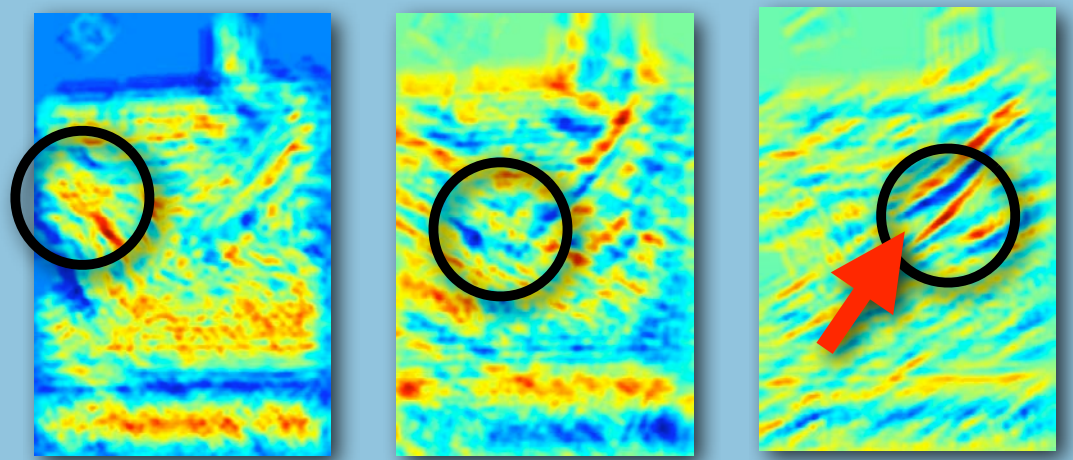
Our part likelihoods





[Ramanan, NIPS'06]





Model Components

TECHNISCHE UNIVERSITÄT DARMSTADT

Appearance Model:



Local Features

orientation 1

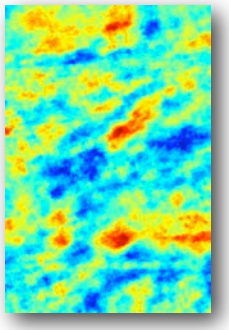


orientation K

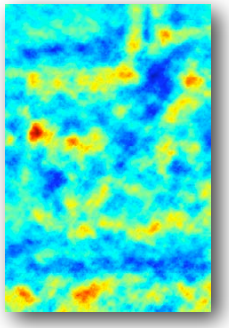


AdaBoost

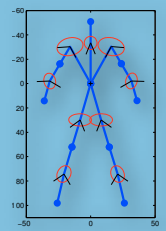
likelihood of part 1



likelihood of part N



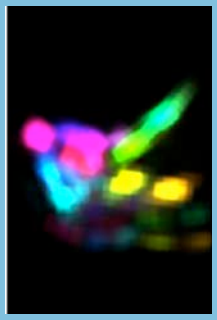
Prior and Inference:



estimated pose



part posteriors

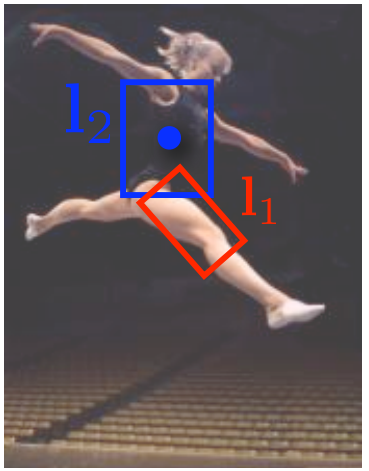


Kinematic Tree Prior

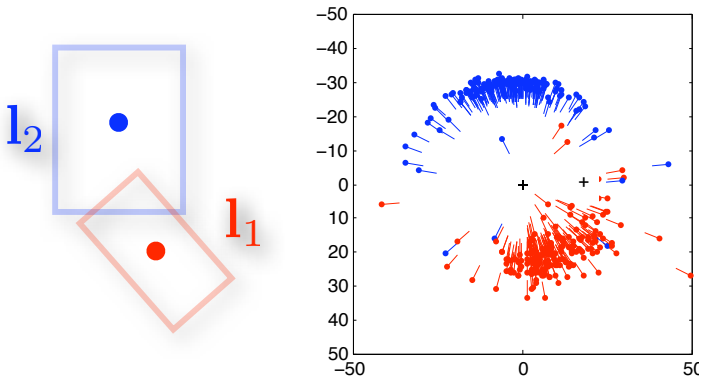
- Represent pairwise part relations [Felzenszwalb & Huttenlocher, IJCV'05]

$$p(L) = p(\mathbf{l}_0) \prod_{(i,j)\in E} p(\mathbf{l}_i|\mathbf{l}_j),$$

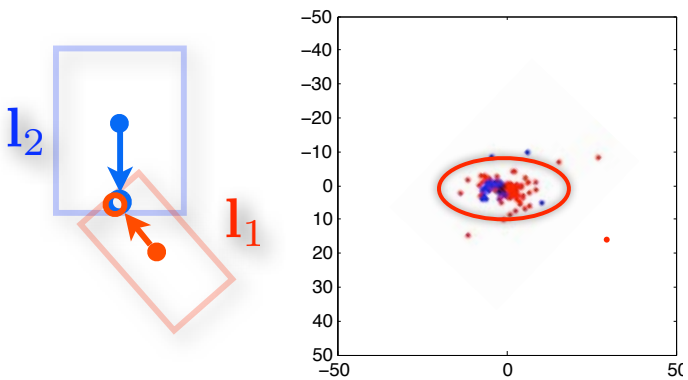
$$p(\mathbf{l}_2|\mathbf{l}_1) = \mathcal{N}(T_{12}(\mathbf{l}_2)|T_{21}(\mathbf{l}_1), \Sigma^{12})$$



part locations relative to the joint



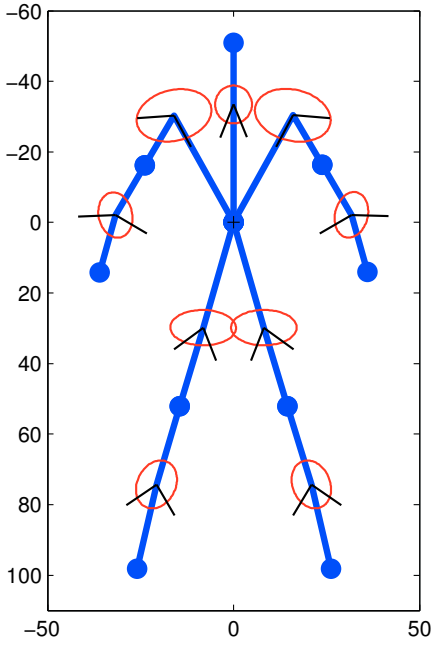
transformed part locations



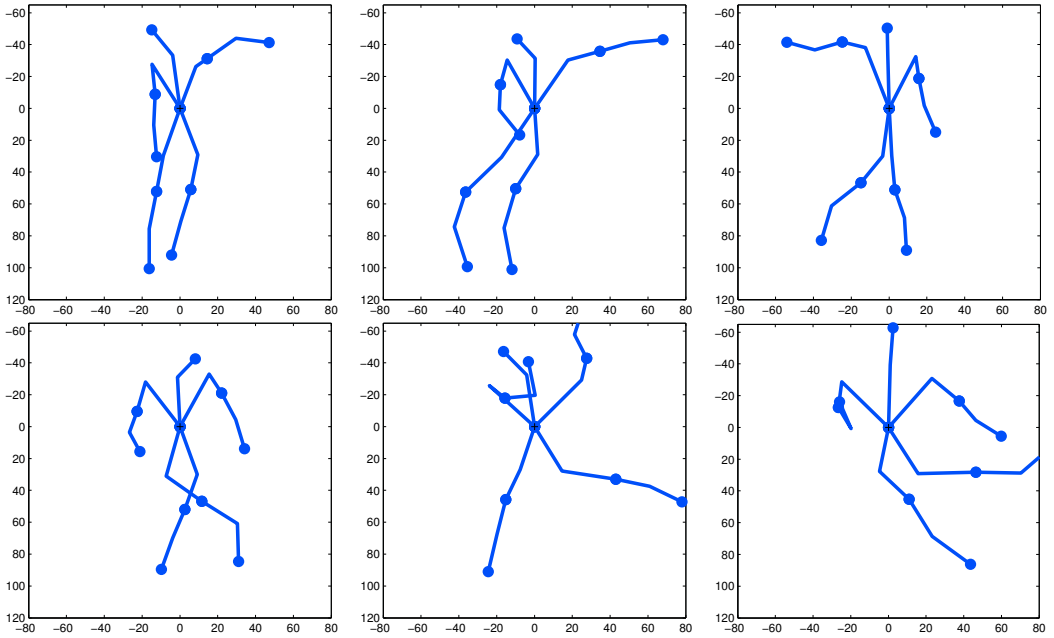
Kinematic Tree Prior

- Prior parameters: $\{T_{ij}, \Sigma^{ij}\}$
- Parameters of the prior are estimated with maximum likelihood

mean pose

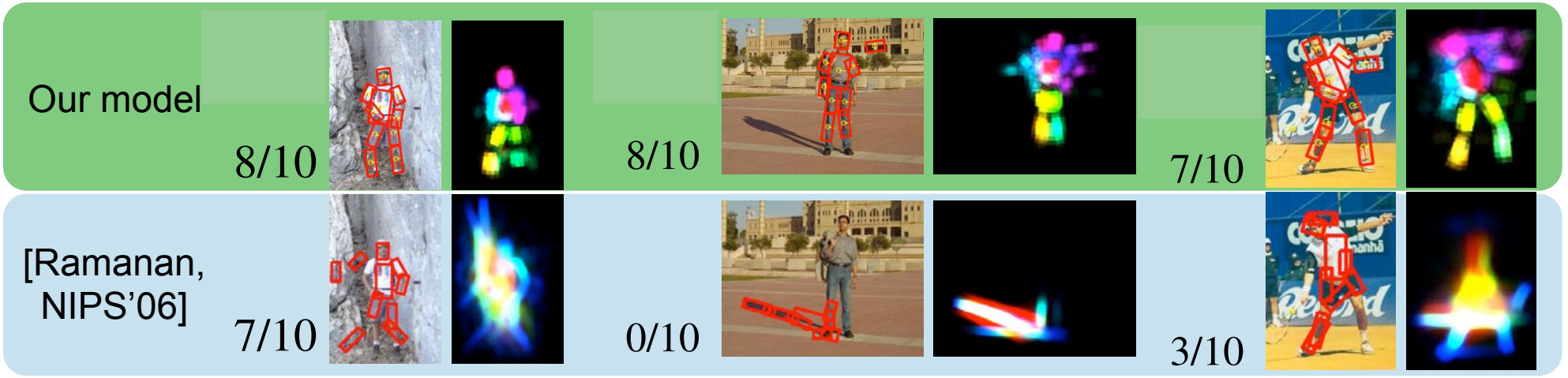


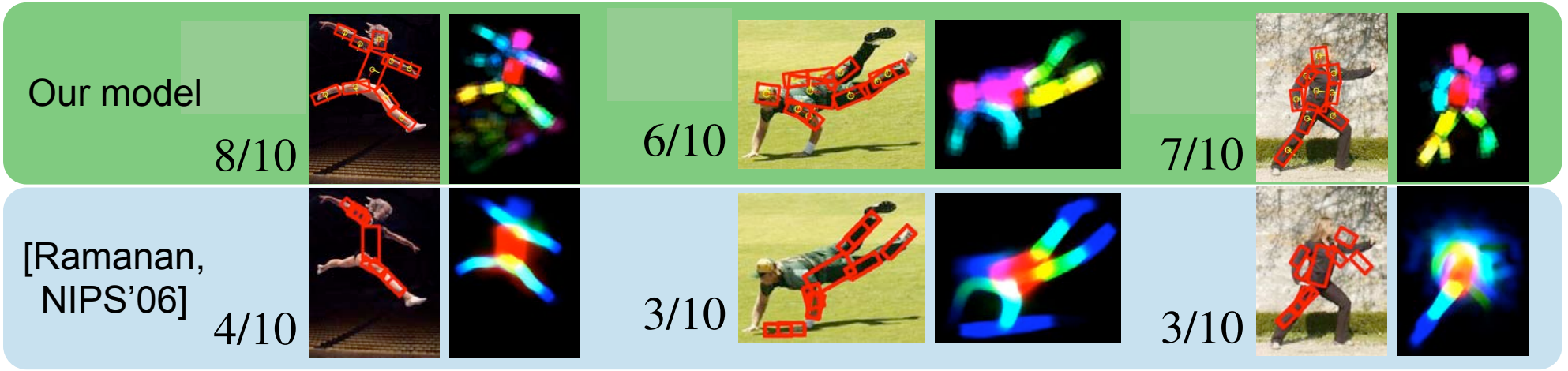
several independent samples



CVPR-09
Results for Articulated Pose Estimation

TECHNISCHE UNIVERSITÄT DARMSTADT

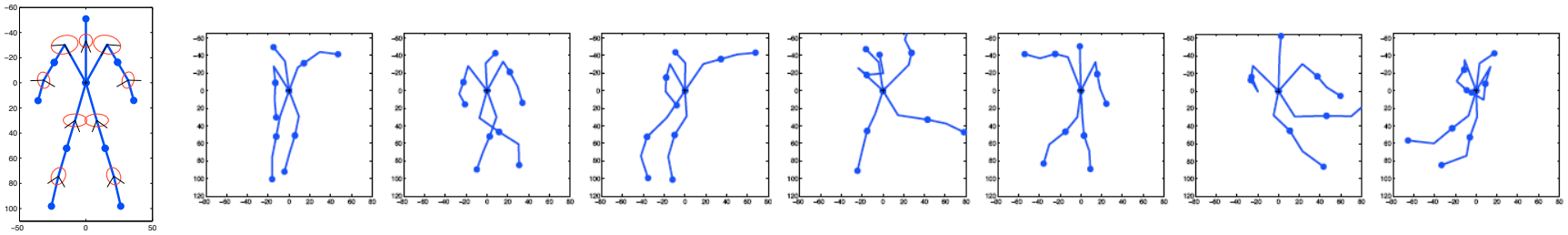




CVPR-09
Results for Articulated Pose Estimation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Generic Kinematic Tree Prior



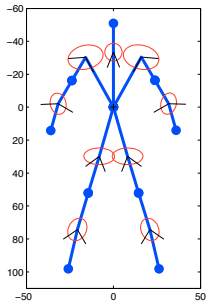
- Quantitative Results

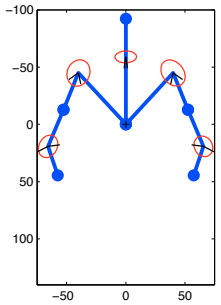
Method	Torso	Upper leg		Lower leg		Upper arm		Forearm		Head	Total
Iterative Image Parsing [13]	48.7	31.2	33.6	26.8	29.7	16.5	15.6	16.5	12.6	34.1	26.5
Our method (detectors only)	29.7	12.6	12.1	20	17	3.4	3.9	6.3	2.4	40.9	14.8
Our method	**82.4**	**69.2**	**62.4**	**65.8**	**46.8**	**47.8**	**45.3**	**32.1**	**30.2**	**74.6**	**55.7**

CVPR-09: Results for Upper Body Detection (Buffy)

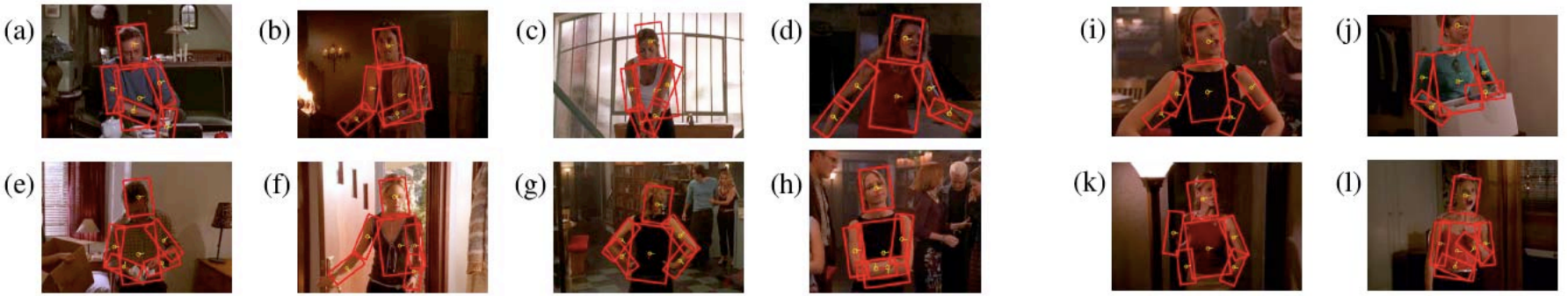
TECHNISCHE UNIVERSITÄT DARMSTADT

- Different Kinematic Tree Priors
  - generic vs. ‘buffy’ (front-back) prior





- Qualitative



- Quantitative

Method	Torso	Upper arm		Forearm		Head	Total
Progressive Search Space Reduction [5]	–	–	–	–	–	–	57.9
Our method (detectors only)	18.9	6.6	7	3.3	2.9	47.2	14.3
Our method (generic prior)	**90.7**	80.2	78.4	40.1	42.3	**95.9**	71.3
Our method (front/back view prior)	**90.7**	**80.6**	**82.1**	**44.2**	**47.9**	95.5	**73.5**

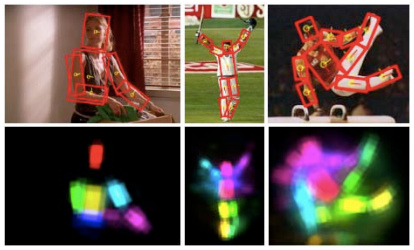
Discussion (people detection)

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Visual People Detection**
  - has made tremendous progress over the last 5+ years
- **Global Feature Based Methods**



  - simple, yet effective to obtain bounding boxes
  - performance starts to saturate (at least for single frame detection)
  - pros: good performance for mid- and large-scale pedestrians, ...
  - issues: small scales (<40 pixels), partial occlusion, requires substantial and representative training data, ...
- **Part-based Models enable Articulated Pose Estimation**
  - more complex models & inference
  - performance can still be improved
  - pros: enable articulated pose estimation & tracking
  - issues: require higher resolution



People Detection: Outlook & Research Directions

TECHNISCHE UNIVERSITÄT DARMSTADT

1. **Integration of Detection & Tracking**
   - from single frame detection,
   - over multi-frame detection
   - to continuous people tracking
2. **Motion Cues as Complementary Information**
   - motion is important cue for humans to detect people
   - many failure cases of single frame people detection seem solvable by using motion
3. **System Integration**
   - integrate detection & tracking with stereo & depth estimation, structure from motion, ground-plane estimation, etc.
   - examples: Gavrila et al.@Daimler & Ess et al.@ETH Zurich

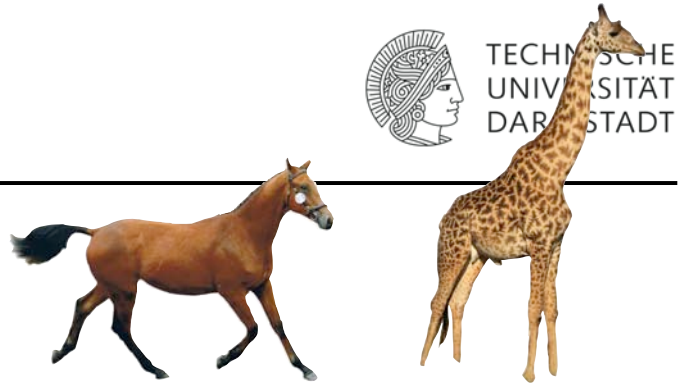
Overview of lecture parts 3 & 4...

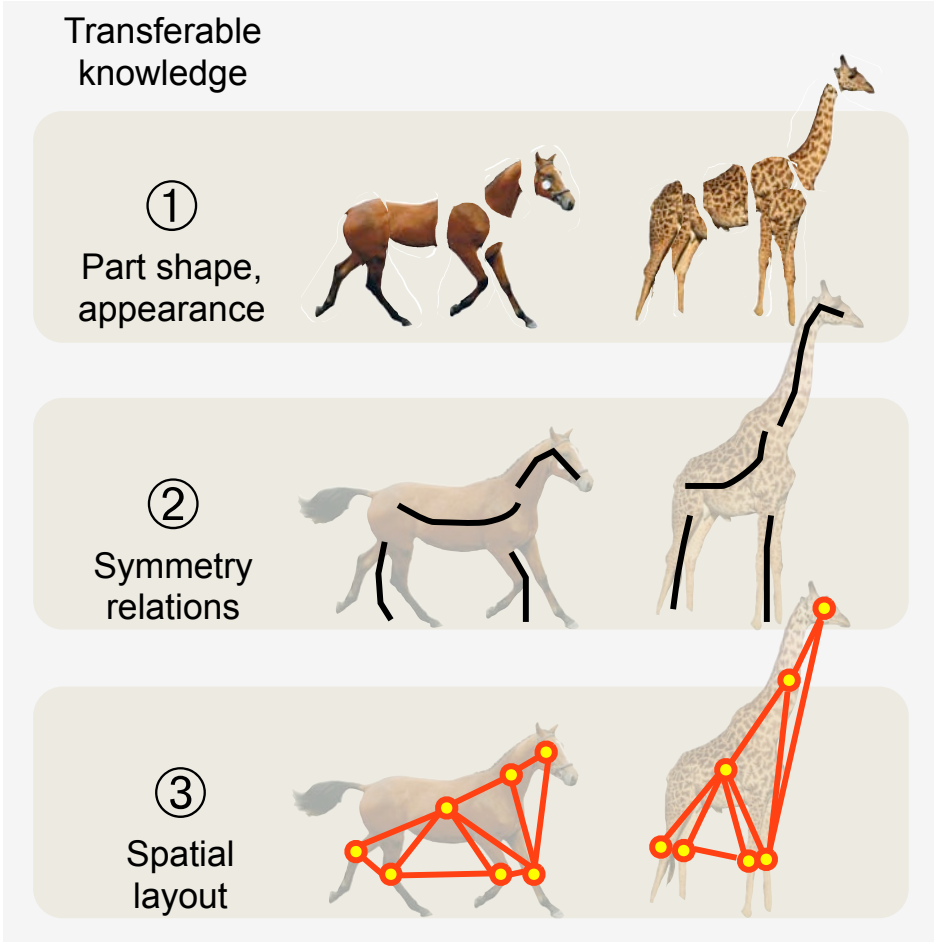
TECHNISCHE UNIVERSITÄT DARMSTADT

- Global Feature Based Methods for People Detection (part 3)
  - A Performance Evaluation of Single and Multi-Feature People Detection [Wojek,Schiele@DAGM-08]
  - Pedestrian Detection: A New Benchmark [Dollar,Wojek,Perona,Schiele@CVPR-09]
  - Multi-Cue Onboard Pedestrian Detection [Wojek,Walk,Schiele@CVPR-09]
- Part-Based Model for People & Object Detection (part 4)
  - Detection by Tracking and Tracking by Detection [Andriluka,Roth,Schiele@CVPR-08]
  - Pictorial Structures Revisited: People Detection and Articulated Pose Estimation [Andriluka,Roth,Schiele@CVPR-09]
  - A Shape-Based Object Class Model for Knowledge Transfer [Stark,Goesele,Schiele@ICCV-09]

Motivation and Contributions

- Motivation
  - Scaling recognition to large numbers of object classes suffers from high amount of needed training data
  - Knowledge transfer from existing models may reduce number of required training examples
- Contributions
  - Novel shape-based object class model for explicit knowledge transfer (①, ②, ③)
  - Effective incorporation of symmetries
  - Competitive shape-based object recognition results

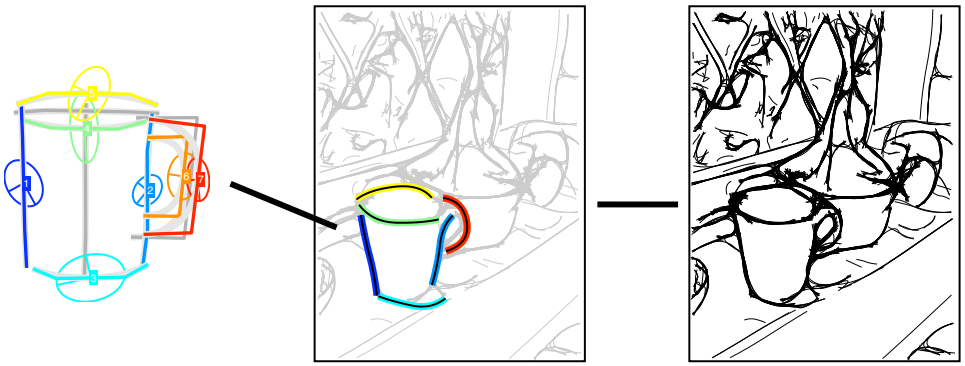




Probabilistic Formulation

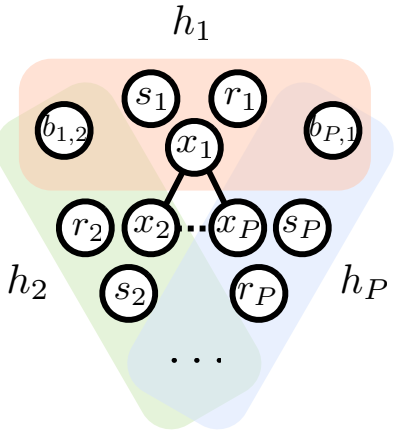
- Hypothesis space
  - Assignments between model parts and local image features [Fergus et al. CVPR'03]

$$H = (h_1, \ldots, h_P)$$

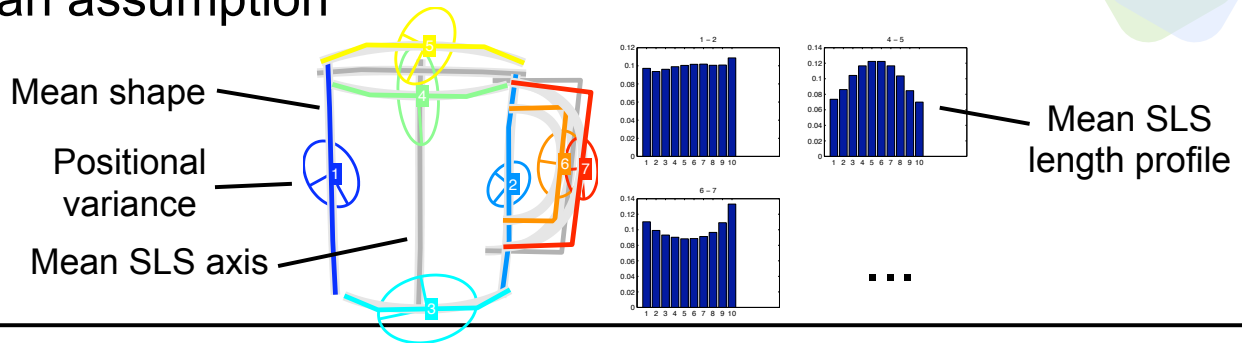


- Factorized joint density

$$p(X, R, S, B, H|\theta) = \underbrace{p(S|H,\theta)}_{Local\ Shape}\ \underbrace{p(B|H,\theta)}_{Symm.\ Rel.}\ \underbrace{p(R|H,\theta)}_{Rel.\ Scale}\ \underbrace{p(X|H,\theta)}_{Layout}\ \underbrace{p(H|\theta)}_{Prior}$$



- Gaussian assumption

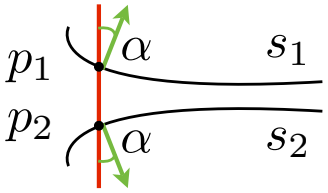


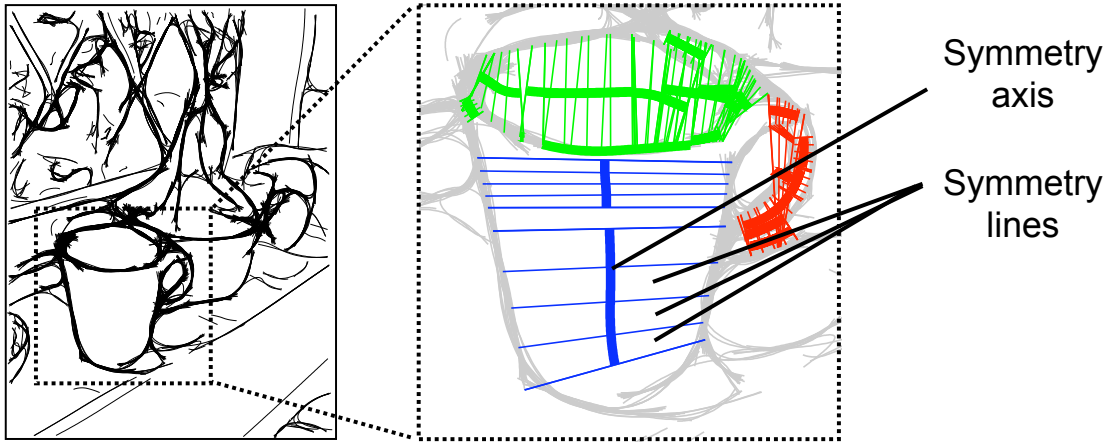
Semi-Local Symmetry Relations

- Smoothed Local Symmetries (SLS)
  - Planar shape analysis [Brady et al. IJRR'84]
  - B-Spline implementation [Saint-Marc at al. PAMI'93]

s_1, s_2
locally symmetric in
p_1, p_2

$\Leftrightarrow$





- Semi-local symmetry descriptor
  - Characterizes a *pair* of B-Spline features
  - Symmetry axis (local shape descriptor)
  - Symmetry lines (PCA on sampled length profile)

Efficient DDMCMC Inference

- MAP hypothesis search
  - Approximated by sampling [Tu et al. IJCV'05] $\hat{H} \sim p(X, R, S, B, H|\theta)$
  - Data-driven part proposals [Lee et al. CVPR'04] $\{q_p(S(h_p)|\theta)\}_{p=1}^{P}$

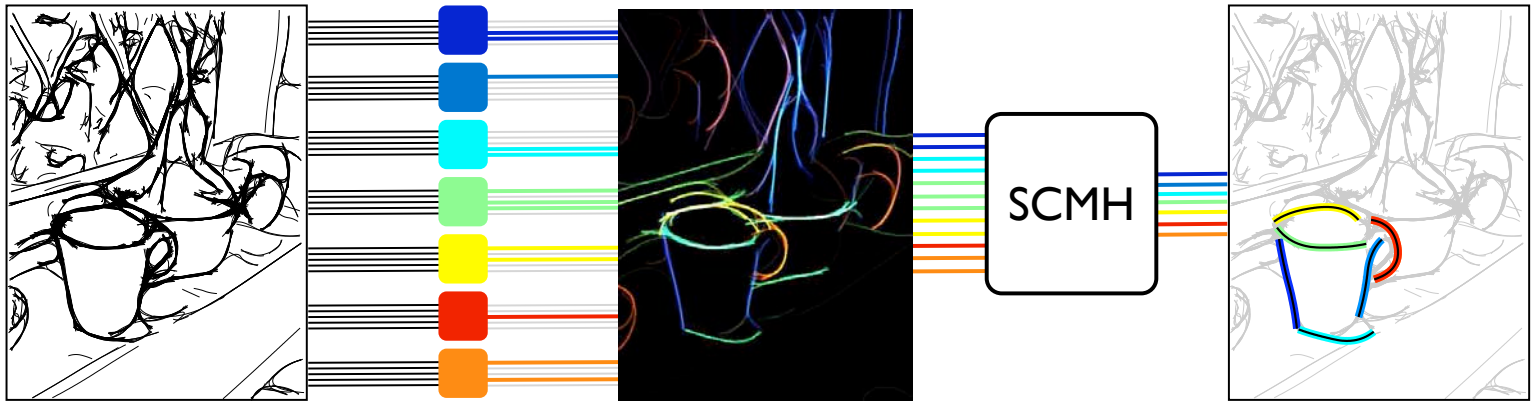


Image features | Part proposals q_p | $\hat{H}$

  - Single-Component Metropolis-Hastings (SCMH) [Gilks et al. '96]

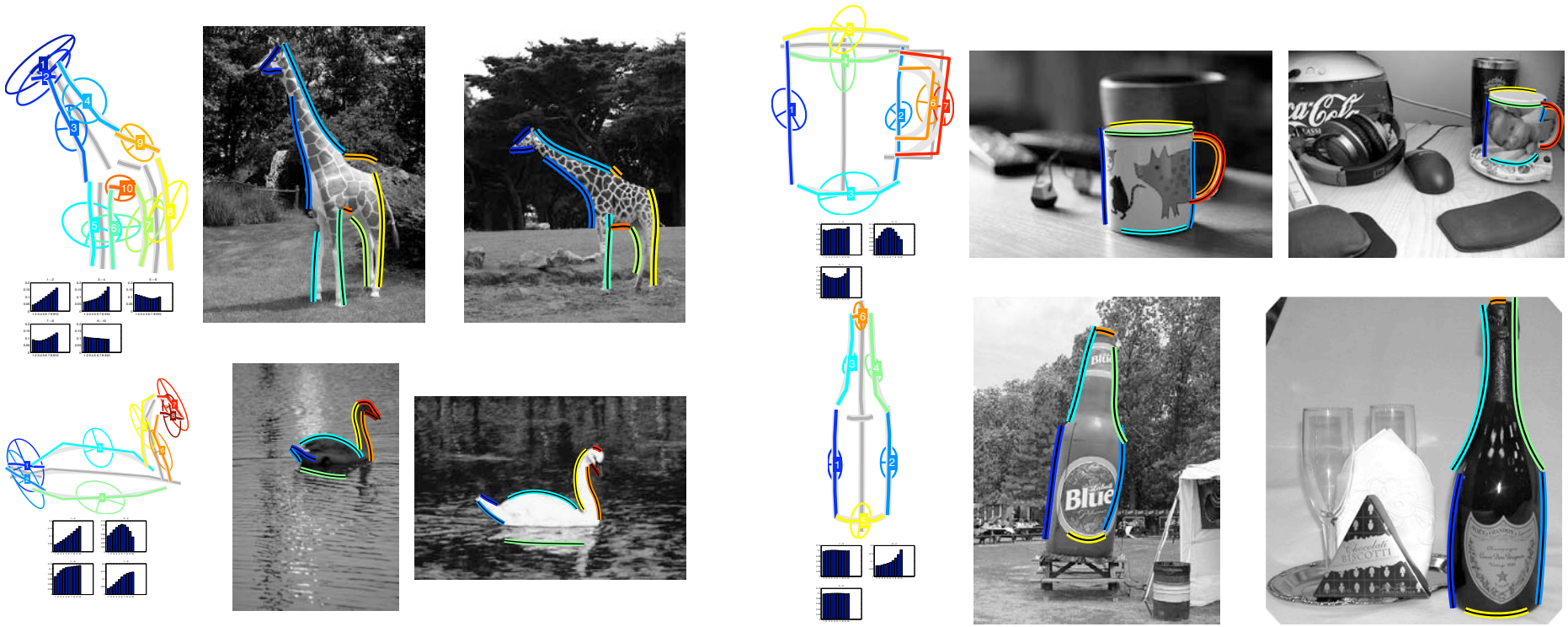
$$H = (H_{-p}, h_p) \xrightarrow{p(accept) = \alpha} H' = (H_{-p}, h'_p)$$

$$\text{where} \quad \alpha = \min\left\{1, \frac{p(X, R, S, B, h'_p|H_{-p}, \theta)\ q_p(S(h_p)|\theta)}{p(X, R, S, B, h_p|H_{-p}, \theta)\ q_p(S(h'_p)|\theta)}\right\}$$

Experiments - ETHZ Shape Classes

- Object classes *Giraffe*, *Swan*, *Mug*, *Bottle*
  - Recall @ 0.4 FPPI, averaged over 5 runs

	[Ferrari et al. CVPR'07]	[Fritz et al. CVPR'08]	Our model	Our model, SLS
average over 4 classes	75.2 (11.0)	83.5 (6.3)	84.3 (5.9)	**89.9 (4.5)**



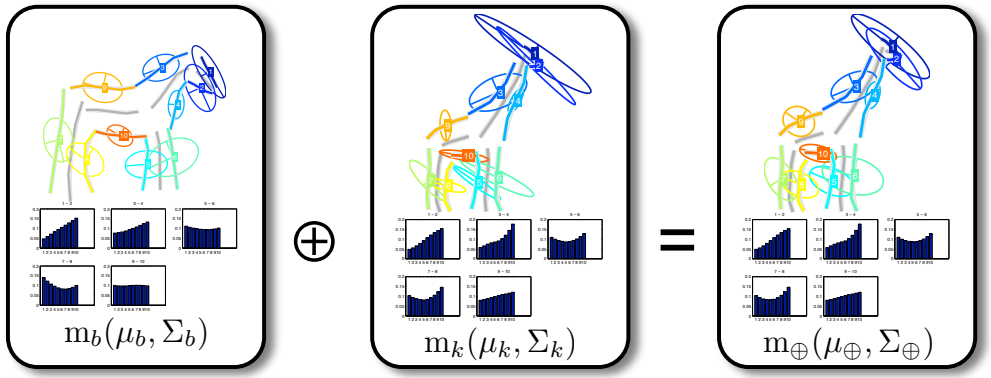
Explicit Knowledge Transfer across Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Full model transfer
  - Direct manipulation of modes/ permissible variations of model densities
  - Inspired by Bayesian model averaging
  - Enabled by functional form of densities (Gaussian)

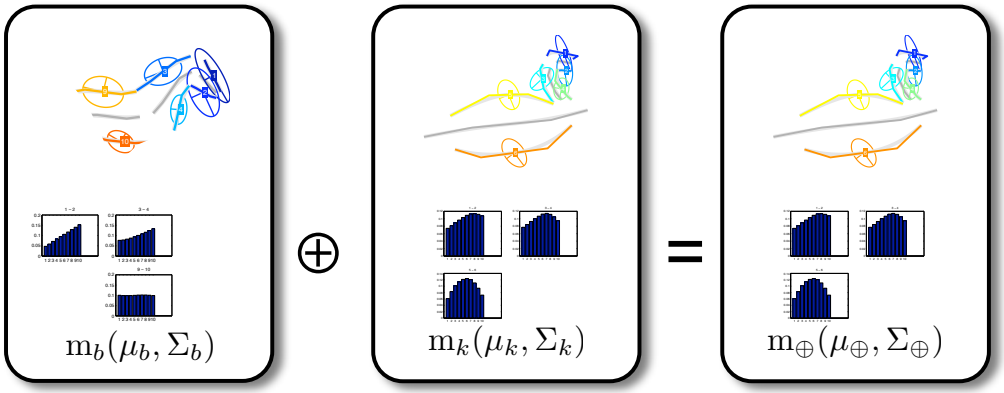
$$\mu_{\oplus} = \mu_k$$
$$\Sigma_{\oplus} = w_b\Sigma_b \; w_k\Sigma_k$$

- Partial model transfer
  - Restriction to relevant model portion
  - Enabled by factorized form of joint density



Base model *(Horse)* | *k*-shot model *(Giraffe)* | base ⊕ *k*-shot model

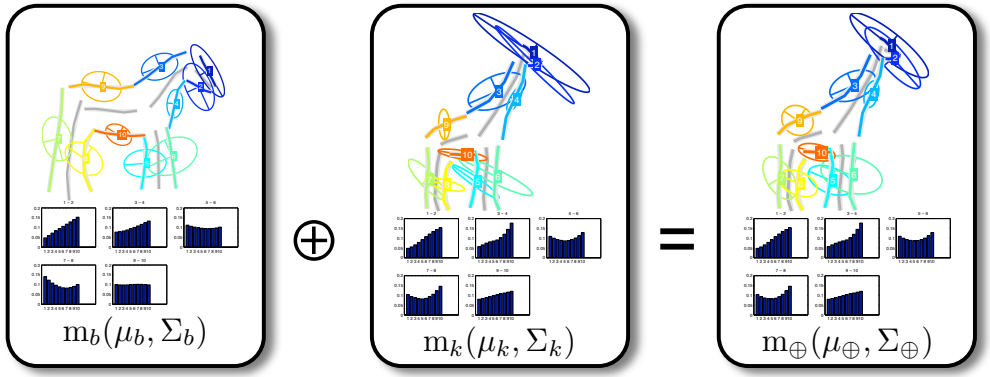
Restriction



Restricted base model *(Horse)* | *k*-shot model *(Swan)* | Restricted base ⊕ *k*-shot model

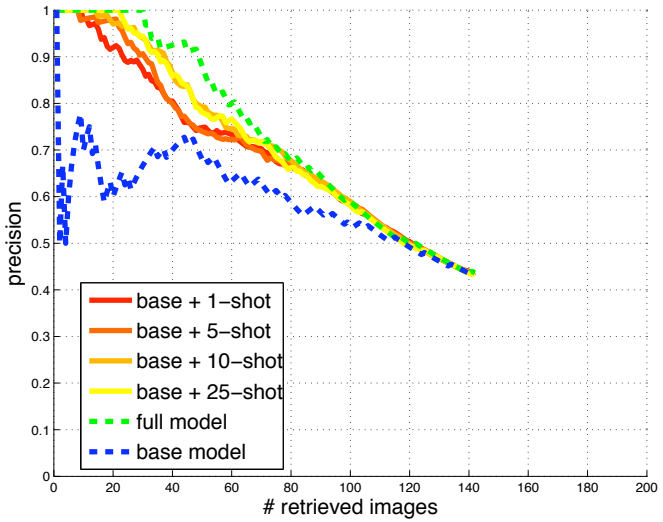
Experiments - Full Model Transfer

- Observations
  - Base model alone *(Horse)* already good on *Elephant*, *Cheetah*
  - Combined base + *k*-shot models often comparable to full models
  - Models including symmetries (SLS) often superior

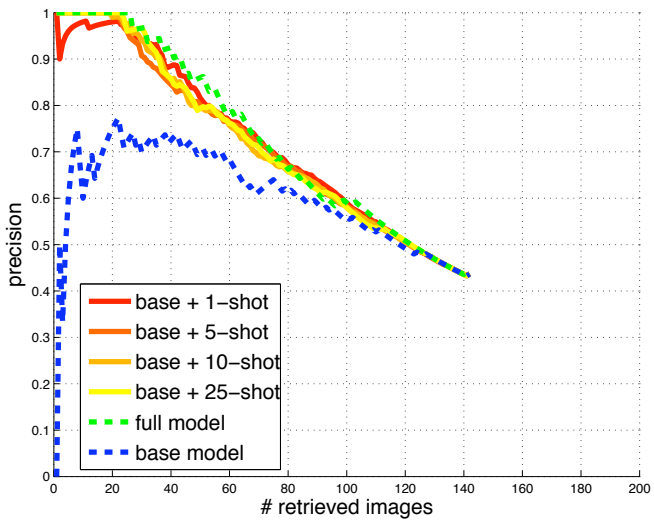


Base model *(Horse)* | *k*-shot model *(Giraffe)* | base ⊕ *k*-shot model

Giraffe

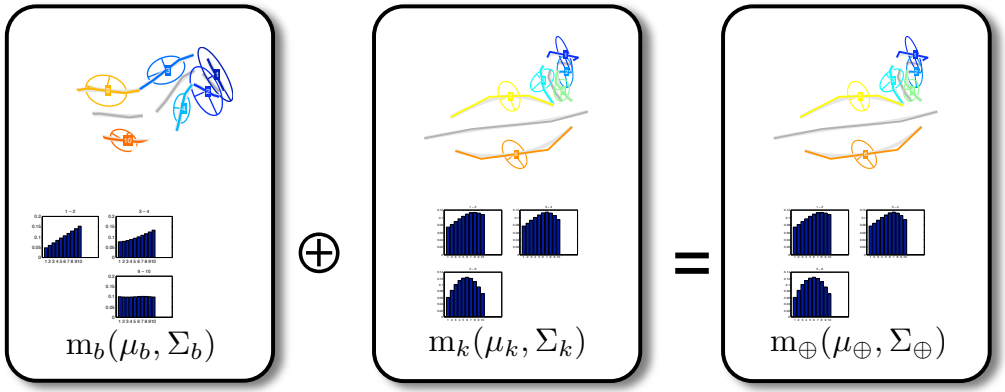


Giraffe, SLS



Experiments - Partial Model Transfer

- Observations
  - Base model *(Horse)* fails on *Swan*
  - Base model *(Mug)* good for *Bottle*
  - Symmetries (SLS) often beneficial

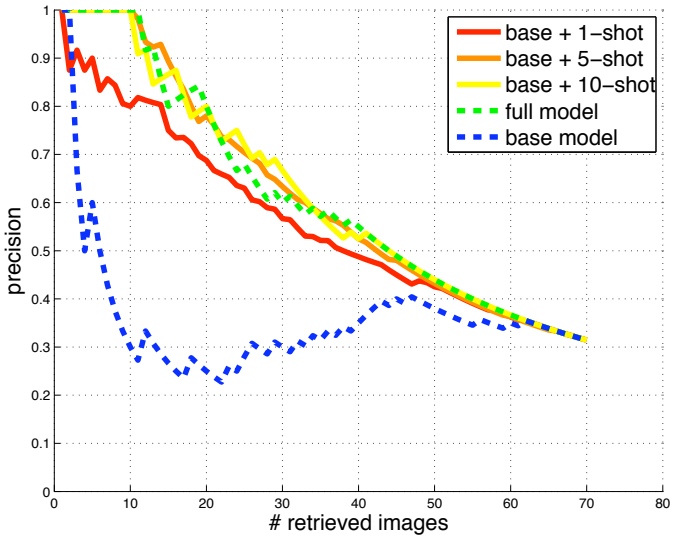


Restricted base model *(Horse)*

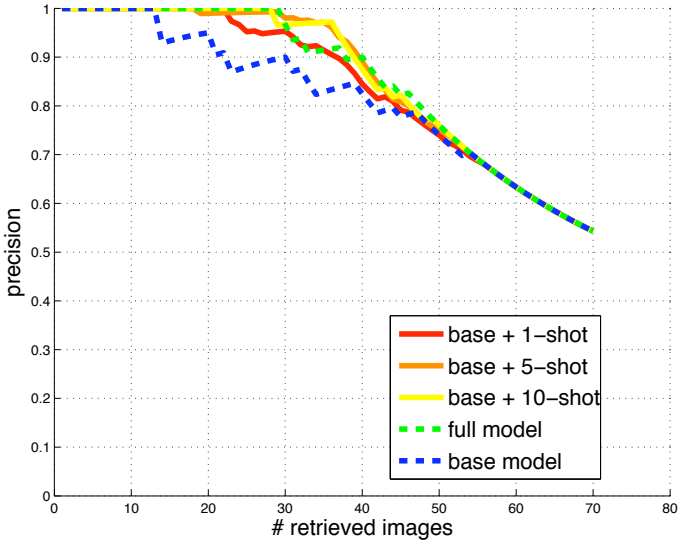
k-shot model *(Swan)*

Restricted base $\oplus$ *k*-shot model

Swan, SLS



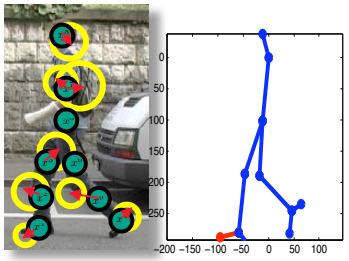
Bottle, SLS

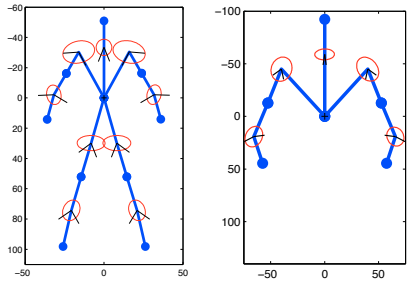


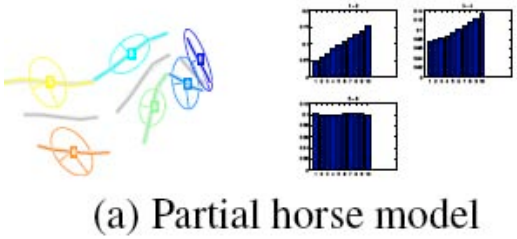
The END

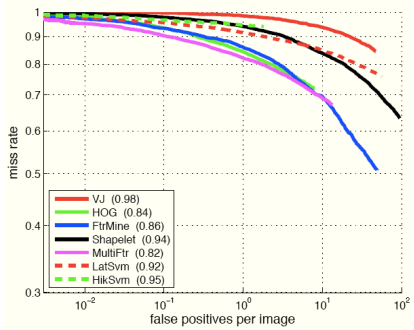
Thank you very much for your attention!



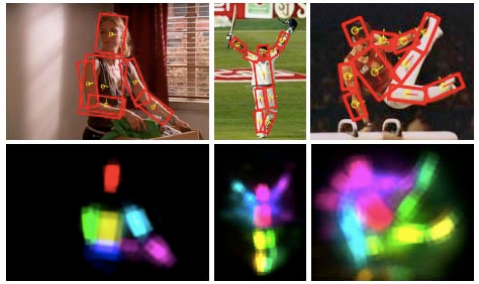


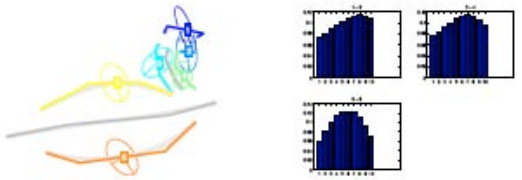












Bernt Schiele
http://www.mis.informatik.tu-darmstadt.de/schiele