Vulnerable Road User

Detection and Orientation Estimation for Context-Aware Automated Driving

Fabian Flohr
University of Amsterdam, Daimler AG (Technical lead for Vulnerable Road User topics)

1

Motives

Vehicle Automation and VRU Safety

Different Reasons for Vehicle Automation

The Disruption Case

Growing on-demand mobility

The Disruption Case Forms of On-Demand Mobility

Car Sharing
You are the driver

+ Flexible, no driver, affordable
- Limited coverage area, parking issues

Taxis / "Ubers"
Chauffeured

+ Versatile, convenient, anywhere
- Costly, needs driver

The Disruption Case Automated On-Demand Mobility

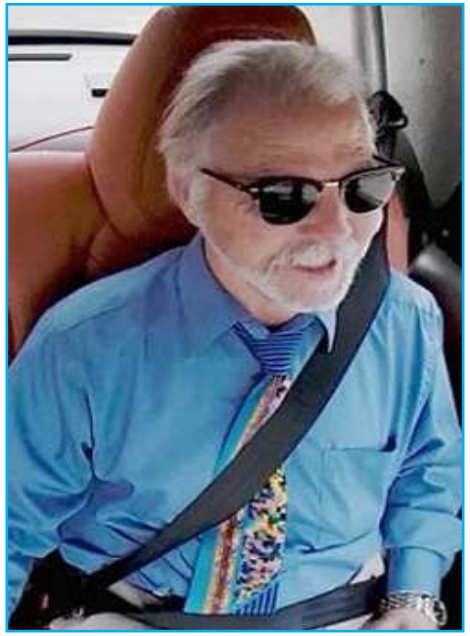
Car Sha You are the	**Robocabs** Self-driving	**/ "Ubers"** uffeured
+ Flexible, no driv - Limited coverage ar	Eventually the best of both worlds?	nvenient, anywhere needs driver

Different Reasons for Vehicle Automation

The Disruption Case

Growing on-demand mobility

The Mobility Case



Mobility for those who cannot drive

The Comfort Case

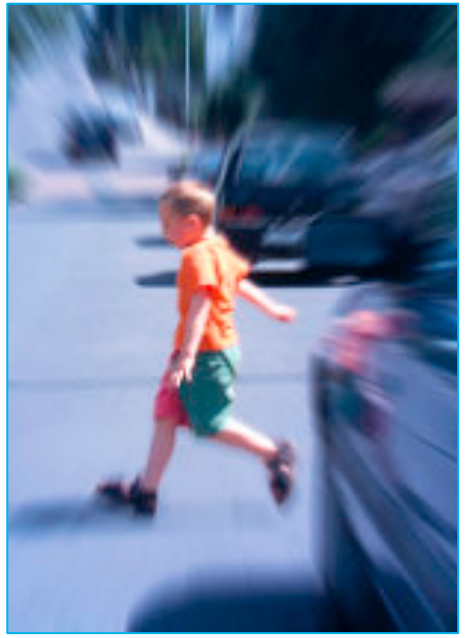
Automation to meet customer demands

The Safety Case

Automation to improve traffic safety

Safety Case Protecting Other Traffic Participants

The Safety Case



Automation to improve traffic safety

1.24 million road traffic deaths world wide

70 VRUs are killed every hour

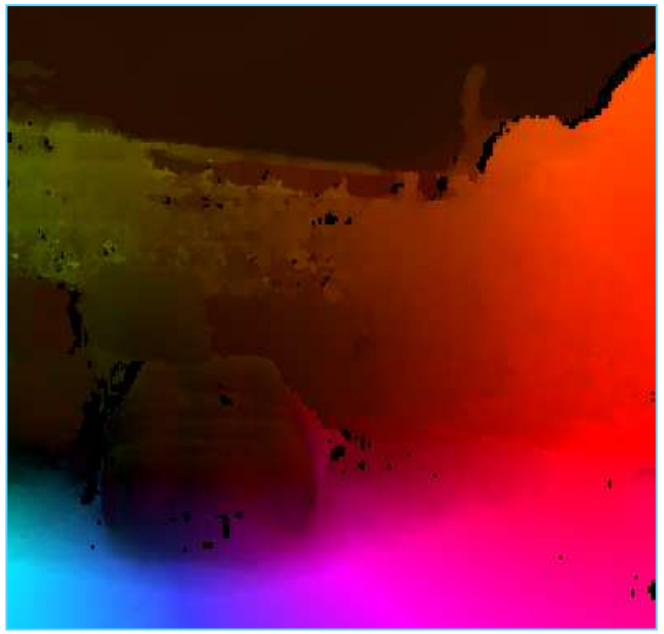
#1 cause of death among those aged **15-29 years**

VRU = pedestrian, cyclist, motorcyclist etc.

Global status report on road safety, WHO 2013

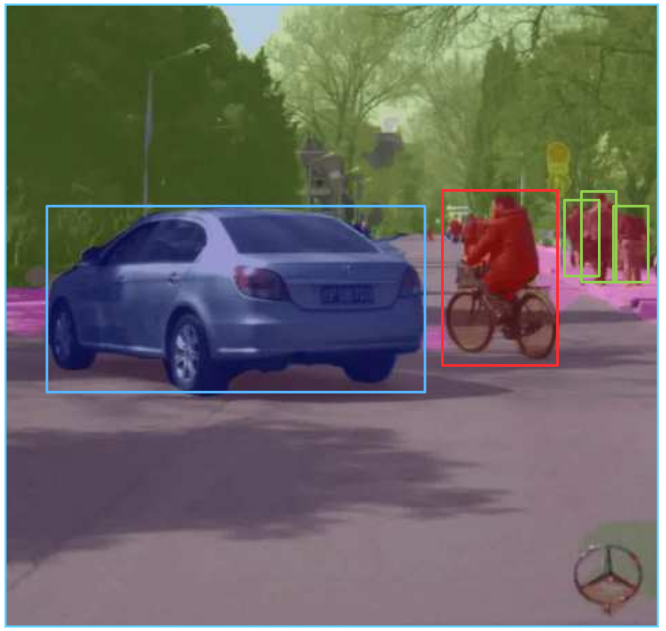
Components Percept and React on Traffic Participants

Dense Spatial Data



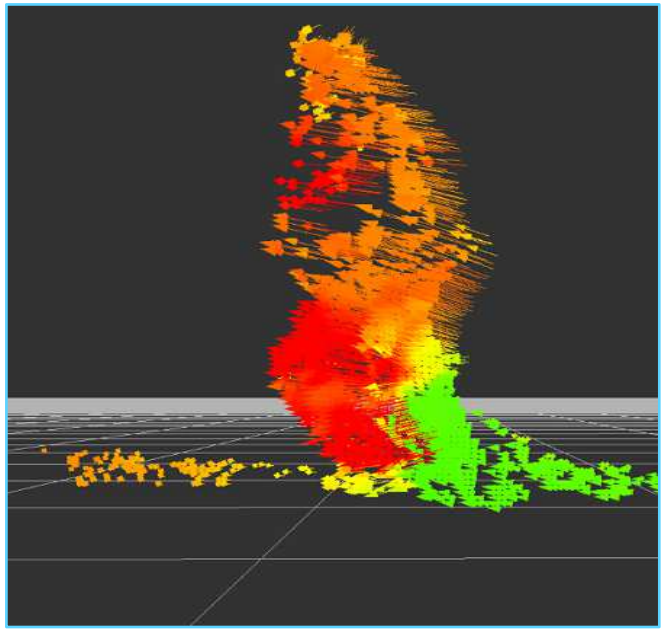
Completeness of data
under all conditions

Object Detection / Segmentation



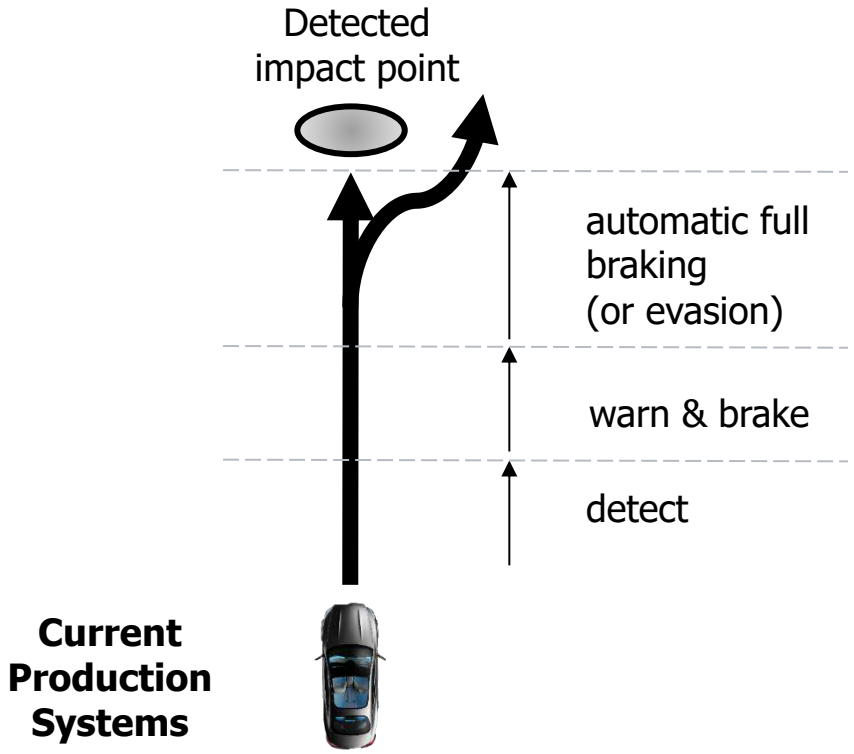
Deep learning for semantic
segmentation and detection

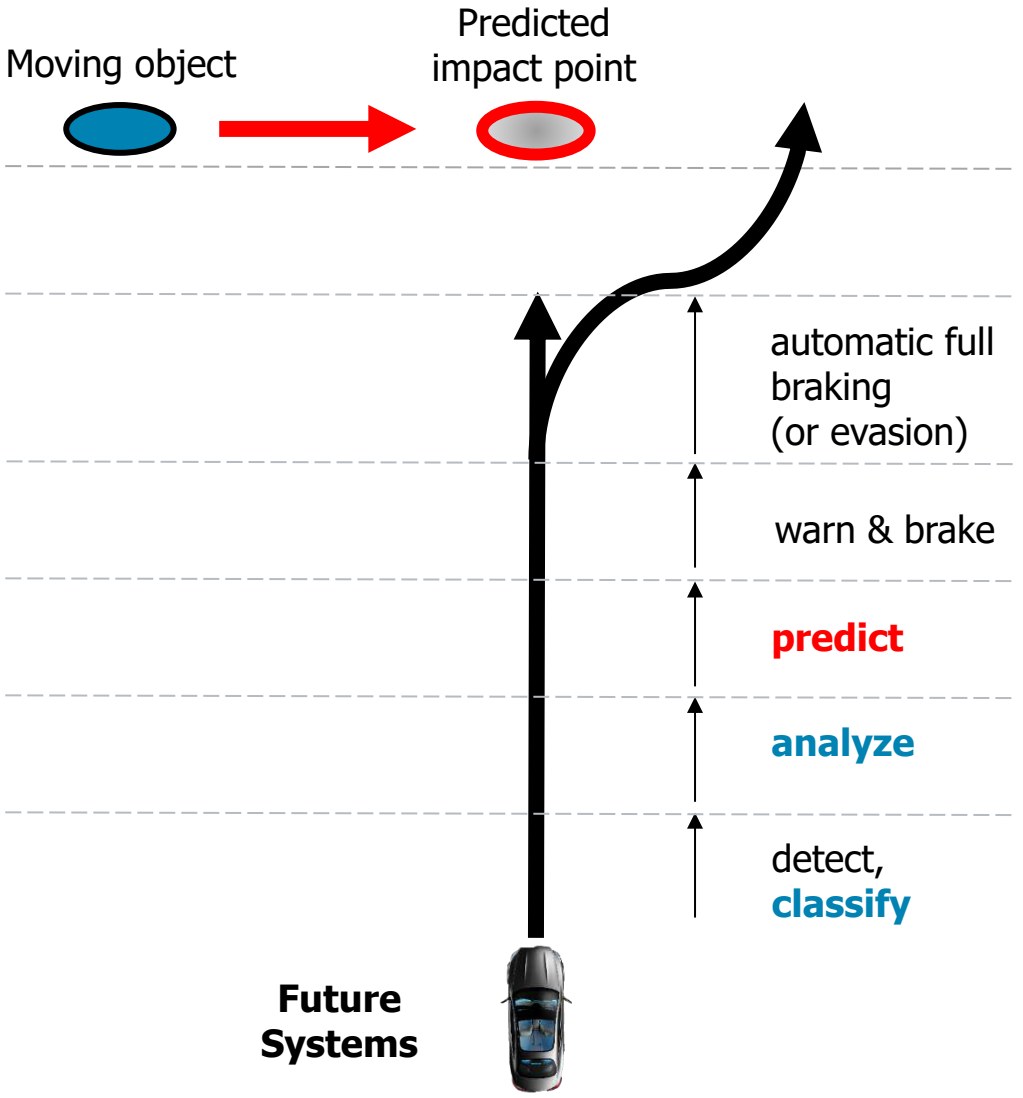
Motion Analysis



Movement calculation
and intention estimation

Safety Case Percept and React on Traffic Participants





Safety Case Improving VRU Path Prediction

Will the pedestrian cross?

A human driver relies heavily on context cues to anticipate how a traffic situation will evolve.

Various context cues influence the pedestrian motion.

J. P. F. Kooij, N. Schneider, F. Flohr and D. M. Gavrila. Context-based pedestrian path prediction. *ECCV,* 2014.

Safety Case Improving VRU Path Prediction

Making use of visual context cues

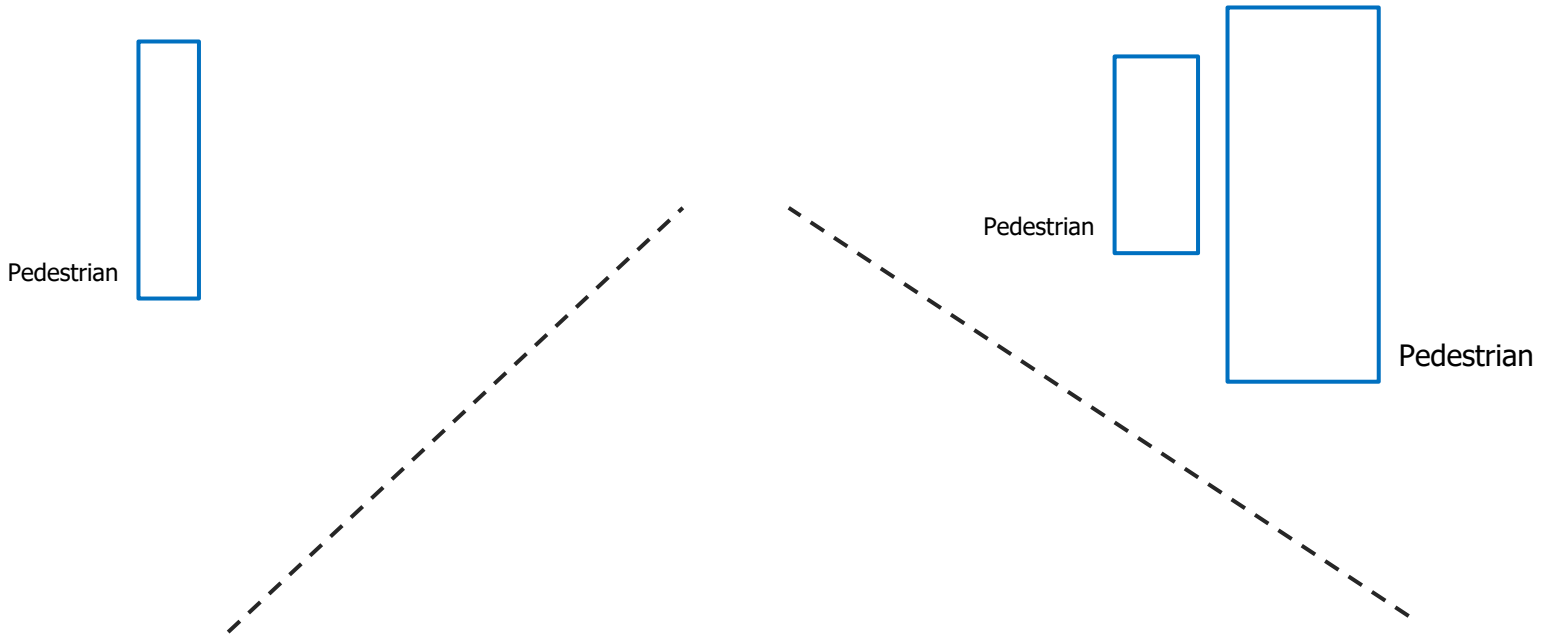
2

A challenge for vehicle automation

Detection of the VRU

Vulnerable Road User Detection

Efficient Cyclist Detection using Stixel-based RoI Proposals within Fast-RCNN architectures



Robust instance-based detection of the VRU as basis for a deeper analysis

X. Li, F. Flohr, Y. Yang, H. Xiong, S. Pan, K. Li and D.M. Gavrila, *„A New Benchmark for Vison-Based Cyclist Detection", IV 2016*

Vulnerable Road User Detection

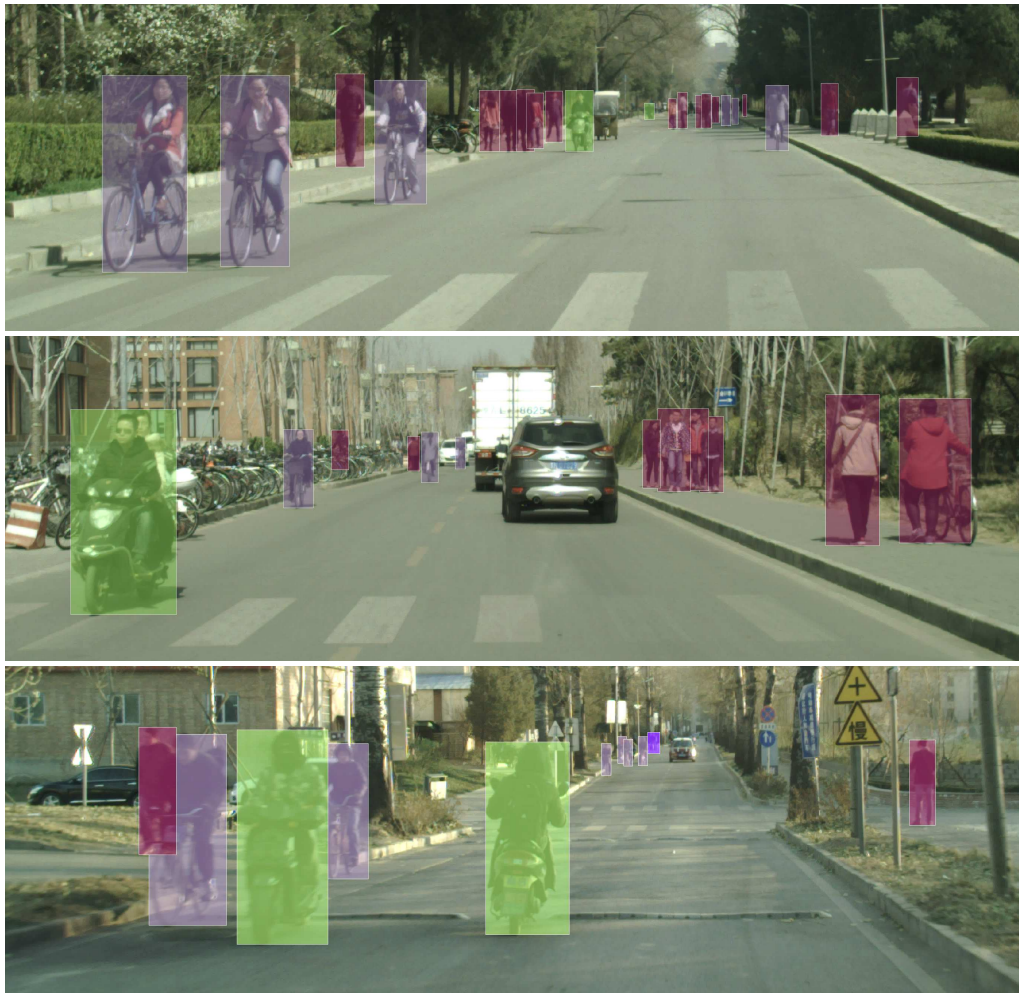
Contributions and Related Work

- New dataset for cyclist detection
  - Sensor recording **vehicle** build-up in China (Stereo camera, Lidar, IMU)
  - Order of **magnitude more cyclists** objects than other automotive datasets [Geiger2013, Zhang2017]
  - **First Chinese automotive dataset** recorded in an urban area with a reasonable size
- Efficient **stixel-based RoI** proposals [Enzweiler2012] used within **Fast-RCNN architectures** [Girshick2015] (**SP-FRCN**)
- **State-of-the-art results** are gained and **compared to**
  - Fast-RCNN when using established **2D proposal** methods [Uijlings2013, Zitnick2014]
  - **Detector ensembles** (as used in [Li2016]) applying handcrafted feature representations [Nam2014, Yang2014]
  - Results show that SP-FRCN outperforms those methods for cyclist detection

A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, *"Vision meets robotics: The KITTI dataset,"* IJRR 2013
S. Zhang, R. Benenson, and B. Schiele, *"CityPersons: A diverse dataset for pedestrian detection,"* CVPR 2017
M. Enzweiler, M. Hummel, D. Pfeiffer, and U. Franke. "Efficient Stixel-based object recognition", IV 2012
R. Girshick, "*Fast R-CNN",* ICCV 2015
J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A.W. M. Smeulders, *"Selective search for object recognition", IJCV 2013*
C. L. Zitnick and P. Dollar, *"Edge boxes: Locating object proposals from edges,"* ECCV 2014
W. Nam, P. Dollar, and J. H. Han, *"Local decorrelation for improved pedestrian detection,"* NIPS 2014
B. Yang, J. Yan, Z. Lei, and S. Z. Li, *"Aggregate channel features for multi-view face detection,"* IJCB 2014

Vulnerable Road User Detection

Tsinghua-Daimler Cyclist Detection Benchmark Dataset



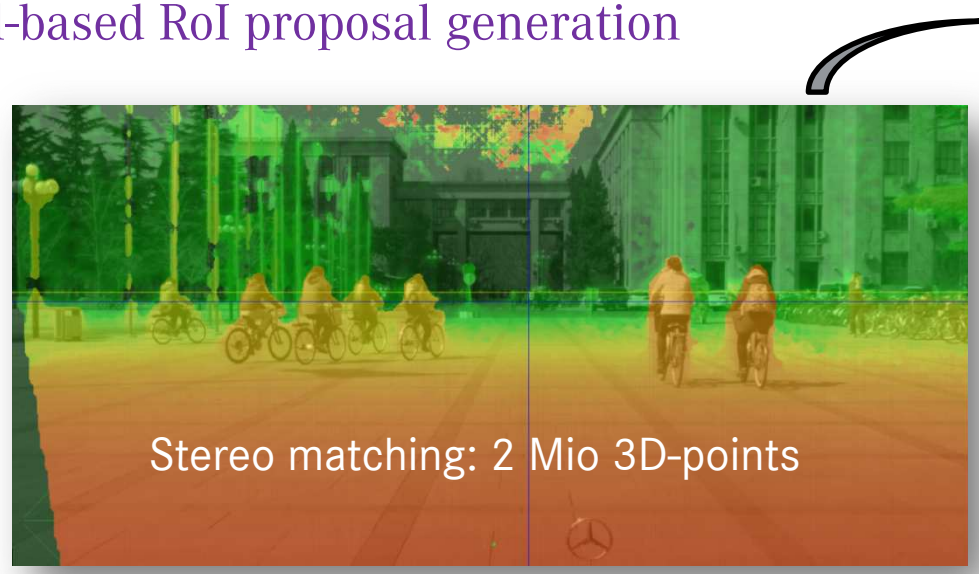
	Training set			Test set	Total
	Set-1a (p.l.)	Set-2 (f.l.)	Set-3 (non-VRU)	Set-4 (f.l.)	
# Images (labeled)	9741	1019	1000	2914	14,674
# VRU Objects	16,202	3016	0	13,143	32,361
# Cyclists	16,202	1301	0	4658	22,161
# Pedestrians	0	1539	0	7380	8919
# Other Riders	0	176	0	1105	1281

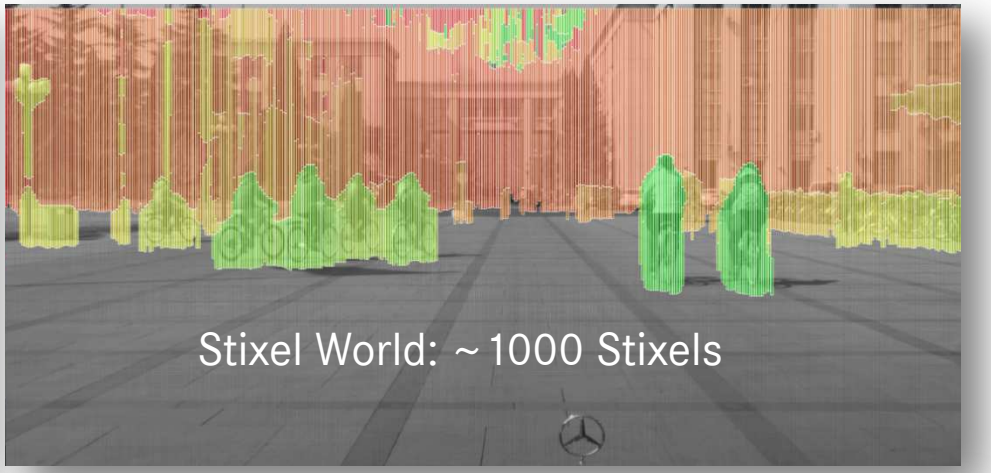
- Includes
  - Left 8-bit color image
  - Disparity information
  - IMU information
  - Bounding box annotations

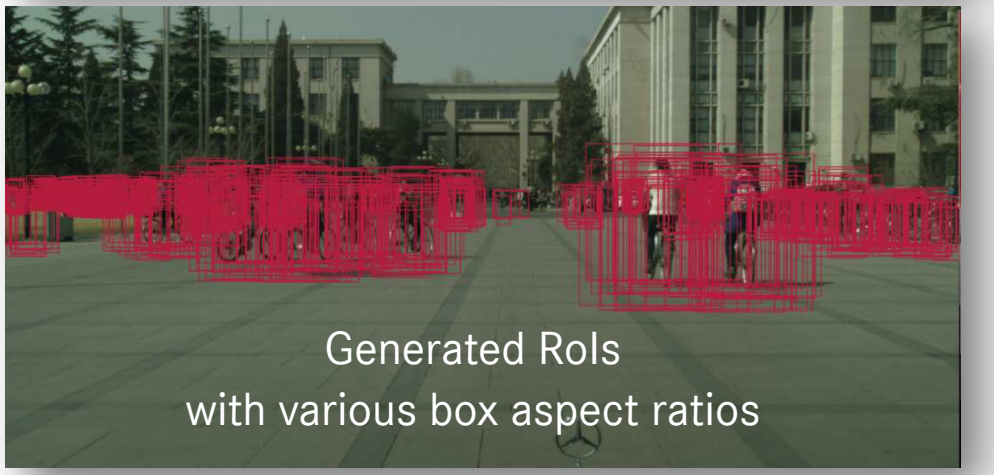
X. Li, F. Flohr, Y. Yang, H. Xiong, S. Pan, K. Li and D.M. Gavrila, *"A New Benchmark for Vison-Based Cyclist Detection", IV 2016*

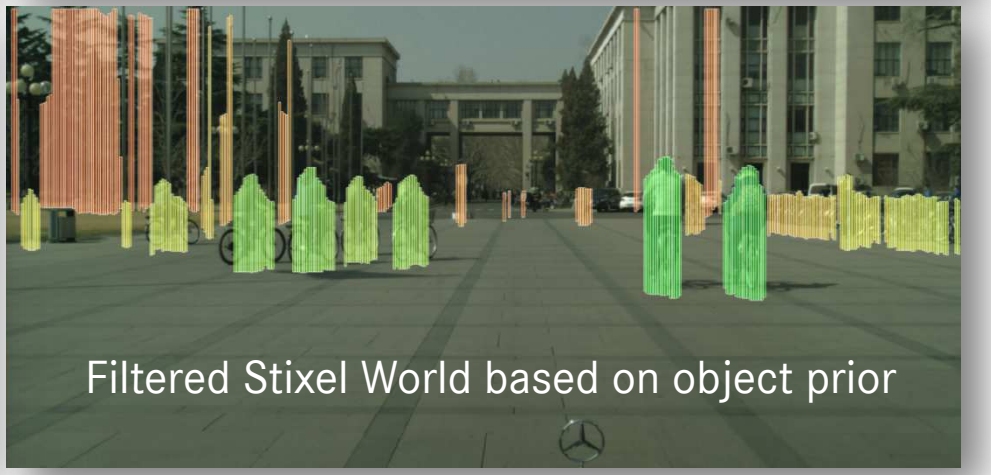
Vulnerable Road User Detection

Stixel-based RoI proposal generation



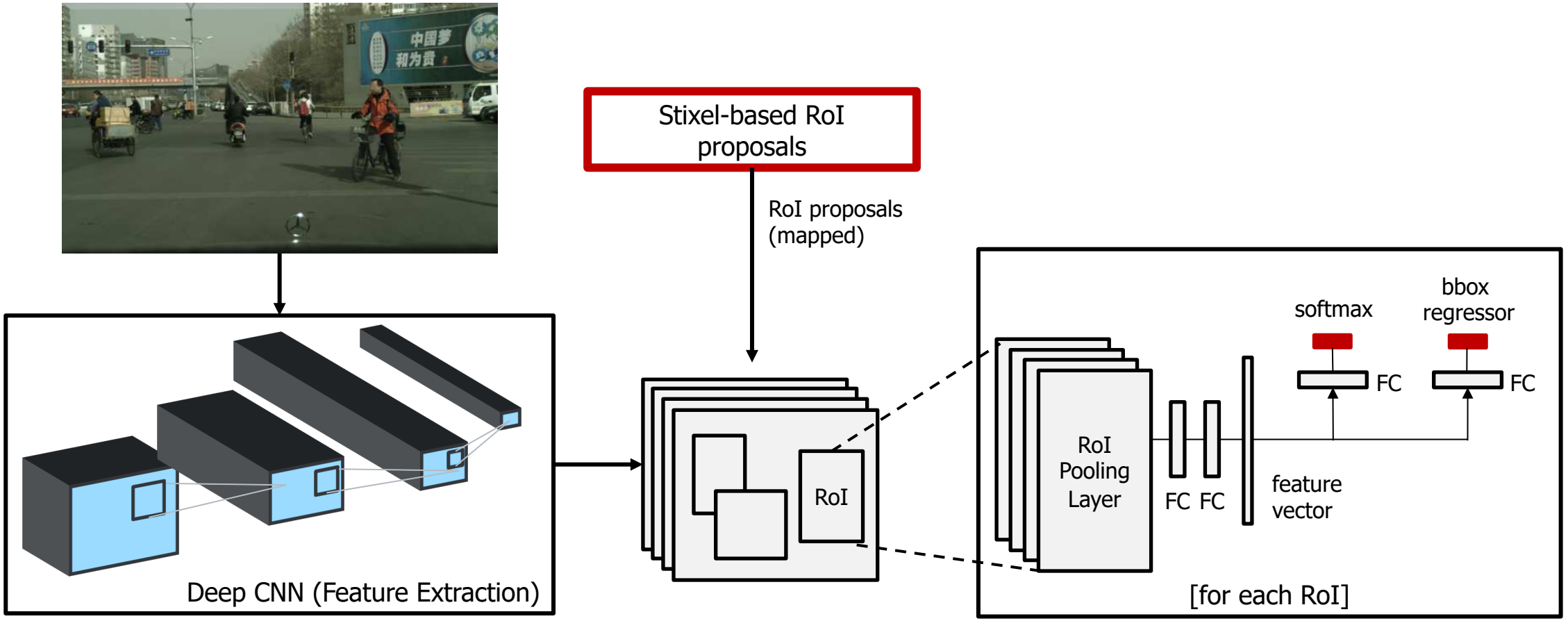






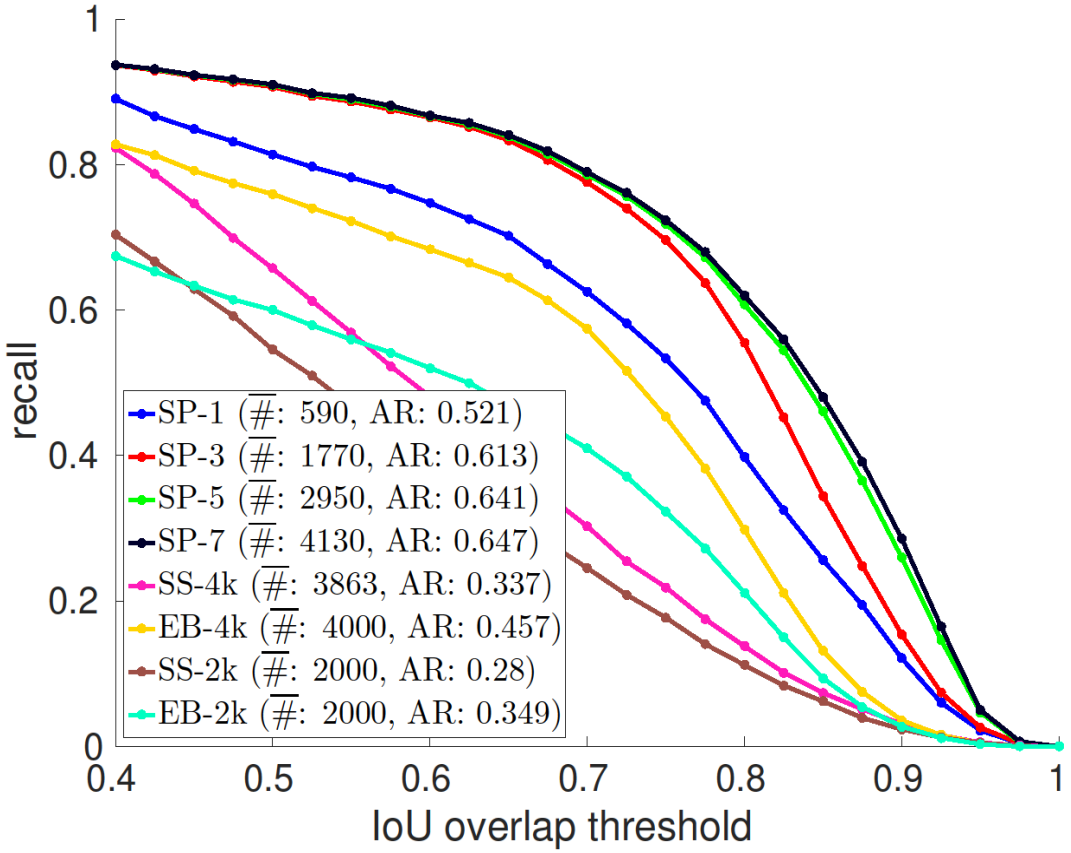
Vulnerable Road User Detection

Efficient Cyclist Detection using Stixel-based RoI Proposals within Fast-RCNN architectures

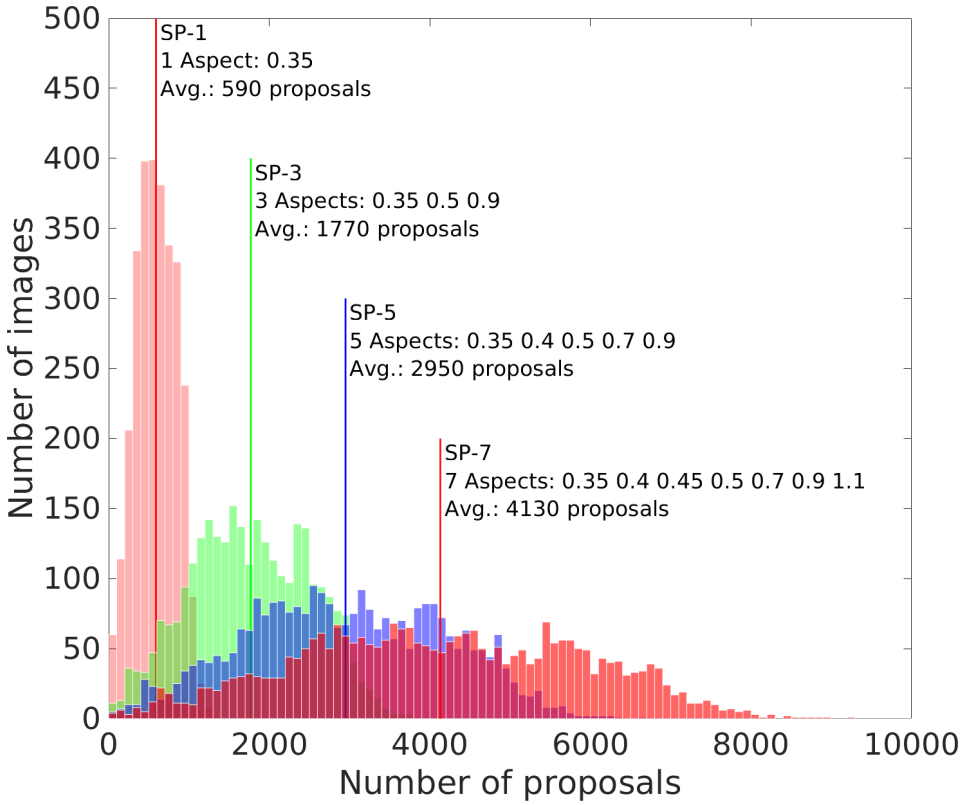


Vulnerable Road User Detection

Proposal Evaluation



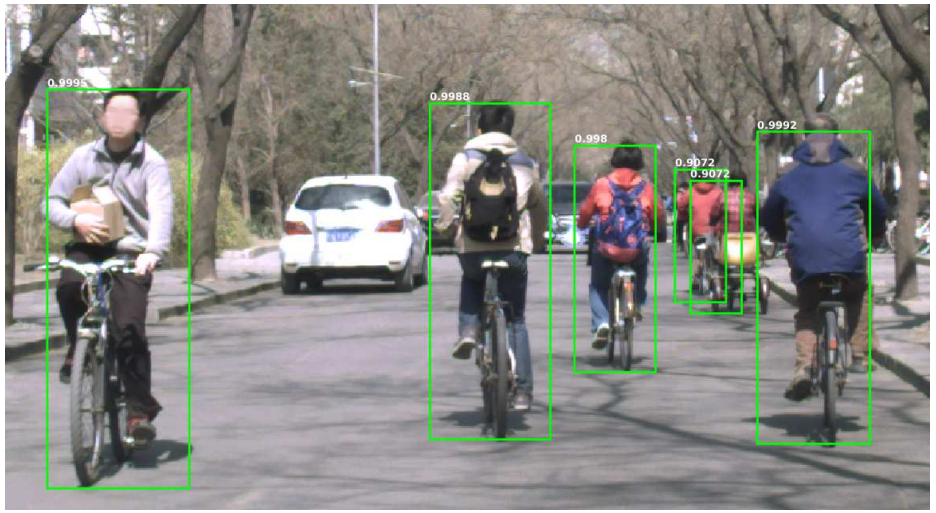
Proposal configurations have been evaluated on validation set.

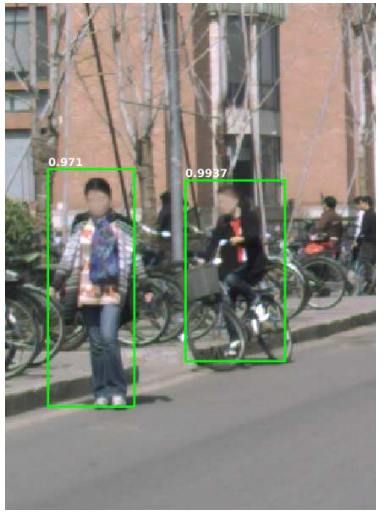


Distribution of the proposal count for different configurations.

Vulnerable Road User Detection

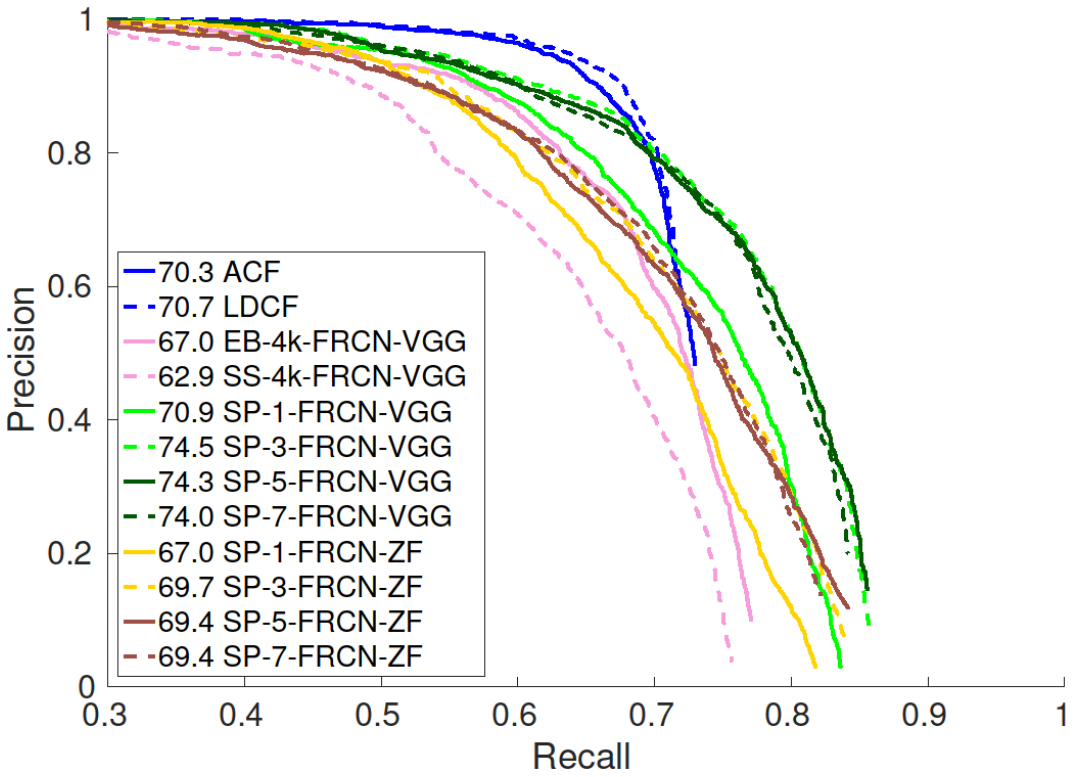
Results





Good detection results

Failure cases



Moderate test set with cyclists with bounding boxes:

- higher than 45 pixels
- occlusion and truncation < 40%

→ SP-X-FRCN with VGG shows higher Average Precision (AP)

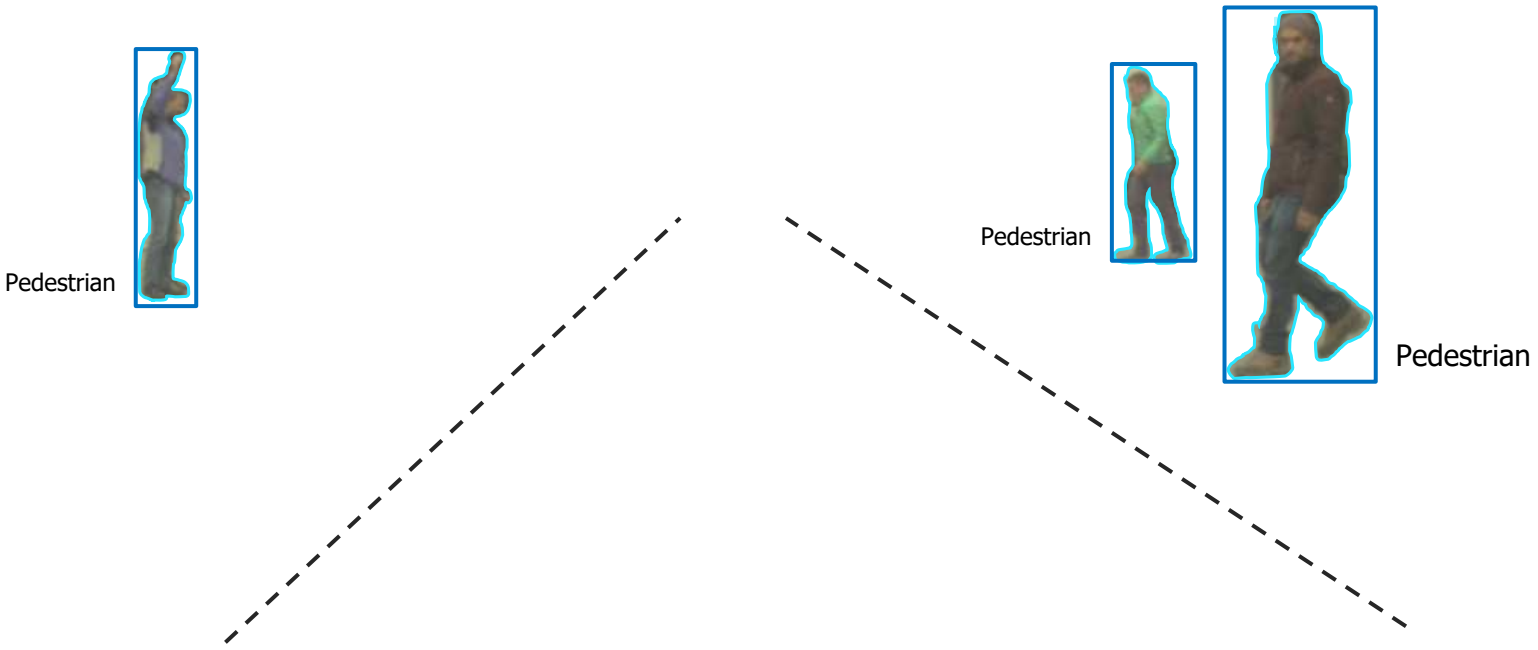
3

A challenge for vehicle automation

Segmentation of the VRU

Vulnerable Road User Segmentation

Accurate Pixelwise Segmentation of the VRU



Segmentation can serve as a basis for extracting further object details, such as part-based orientations or posture details

F. Flohr, D. Gavrila. *"PedCut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues"*, BMVC 2013

Vulnerable Road User Segmentation

Contributions and Related Work

- Combination of **top-down shape information** with **bottom-up data cues** to cope with the large variation of pedestrian appearances across cluttered backgrounds.
- **Multiple Statistical Shape Models** (MSSMs) [Giebel2002] are used to cope with stronger shape aspect variations.
- EM-like refinement within a using a **Conditional Random Field** (CRF) formulation.
  - Compared to [Kokkinos2009, Kumar2010], multiple data cues (color, disparity and a discriminative superpixel classifier) are combined and MSSMs are used.
- Results are presented on the public Penn-Fudan pedestrian dataset [Wang2007], where our method **outperforms the state-of-the-art** [Bo2011, Eslami2012].
- A **new pedestrian segmentation dataset** with disparity information is introduced.

J. Giebel and D. M. Gavrila, *"Multimodal shape tracking with point distribution models,"* DAGM 2002
I. Kokkinos and P. Maragos, *"Synergy between object recognition and image segmentation using the expectation-maximization algorithm"*, PAMI 2009
M. P. Kumar, P. H. S. Torr, and A. Zisserman, "Objcut: Efficient segmentation using top-down and bottom-up cues," *PAMI 2010*
L. Wang, J. Shi, G. Song, and I.-F. Shen, *"Object detection combining recognition and segmentation"*, ACCV 2007
Y. Bo and C. C. Fowlkes, *"Shape-based pedestrian parsing", CVPR 2011*
S. M. A. Eslami and C. Williams, *"A generative model for parts-based object segmentation"*, NIPS 2012

Vulnerable Road User Segmentation

Accurate Pixelwise Segmentation of the VRU

Chamfer template matching [Gavrila2007]

Shape initialization (template matching)

Multiple Statistical Shape Models

MSSM

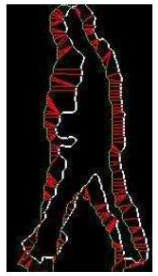
selected SSM

SSM shape

SSM fitting (M-Step)

CRF segmentation (E-Step)

segmentation



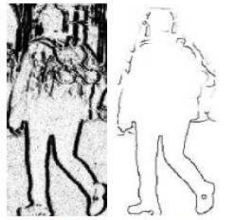
Active Shape Matching (ASM) [Cootes1995]

unary terms

E_u^{BDT} E_u^{SP} E_u^{CP} E_u^{DP}

BDT: Superpixel classification [Hoiem2007]
SP: Shape (distance transform) [Kumar2010]
CP: Color (Gaussian Mixture)
DP: Disparity (Gaussian Mixture)

pairwise terms



E_p^L E_p^C

L: Color-based
C: Contour-based

D. M. Gavrila, "A Bayesian, exemplar-based approach to hierarchical shape matching," *PAMI 2007*
T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, "Active shape models-their training and application," *CVIU 1995*
D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *IJCV 2007*
M. P. Kumar, P. H. S. Torr, and A. Zisserman, "Objcut: Efficient segmentation using top-down and bottom-up cues," *PAMI 2010*

Vulnerable Road User Segmentation

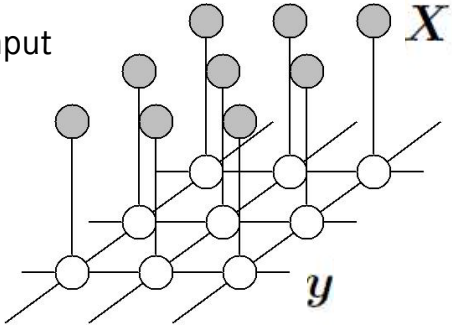
CRF Segmentation

Minimize the energy: $\boldsymbol{y}^* = \underset{\boldsymbol{y}\in\boldsymbol{\mathcal{L}}}{\operatorname{argmin}}\, E(\boldsymbol{y}|\boldsymbol{X})$

$$E(\boldsymbol{y}|\boldsymbol{X}) = \sum_{i\in\mathcal{V}} E_u(y_i|\boldsymbol{X}) + \sum_{i,j\in\mathcal{E}} E_p(y_i, y_j|\boldsymbol{X})$$

$\boldsymbol{X}$ Data input

$\boldsymbol{y}$ Labels

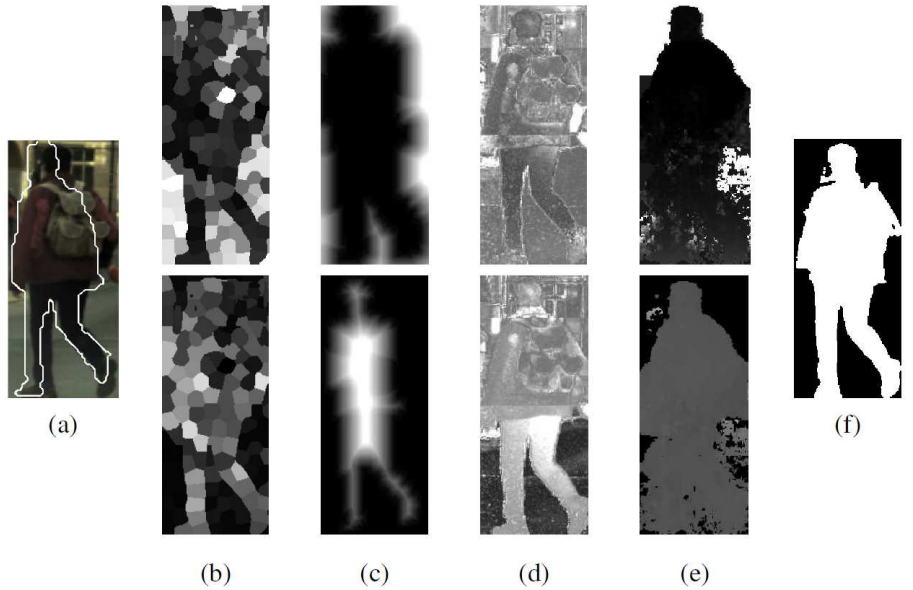


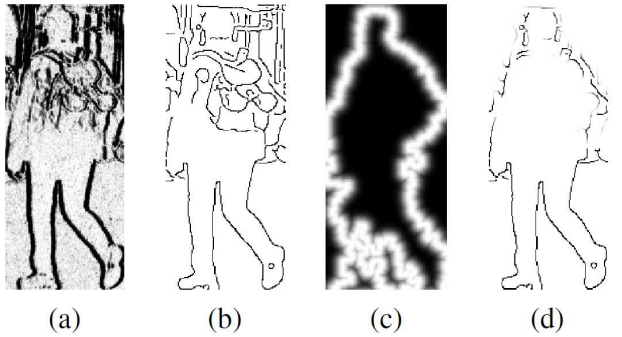
Unary potentials (u)

- **Superpixel** classification based on Boosted Decision Tree ensemble (b) (dSIFT, Texton features) [Hoiem2007].
- **Shape** potential (c) calculated by a distance transformation obtained from the shape contour [Kumar2010].
- GMM-based **color/disparity** potential (d/e);

Pairwise potentials (p)

- **Color-sensitiv** potential (a) [Boykov2001].
- **Contour-sensitive** potential (b), weighted based on the disparity information (c→d).





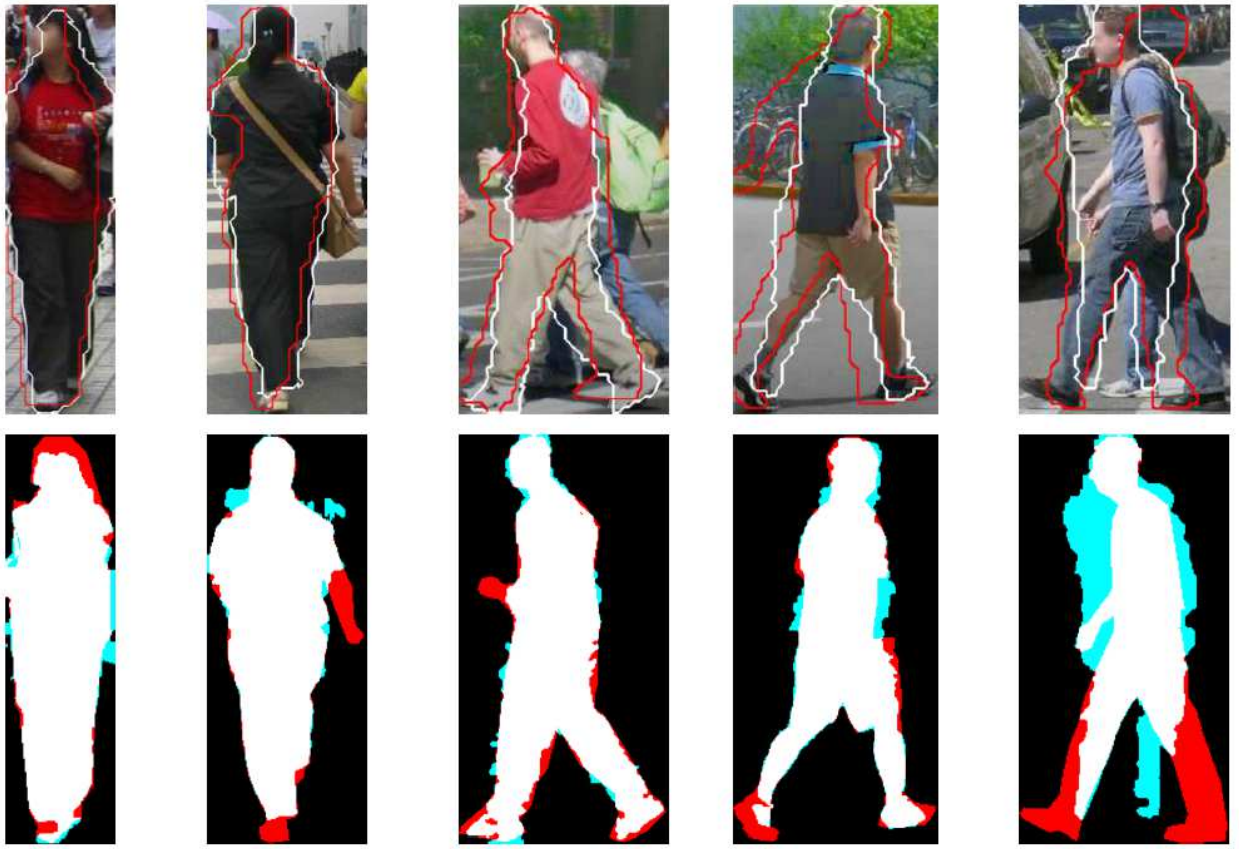
D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *IJCV 2007*

M. P. Kumar, P. H. S. Torr, and A. Zisserman, "Objcut: Efficient segmentation using top-down and bottom-up cues," *PAMI 2010*

Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," ICCV 2001

Vulnerable Road User Segmentation

Results on the Penn-Fudan Dataset



Results on the
Penn-Fudan dataset.

	Foreground	Background	Average
SP	67.2%	71.6%	69.4%
SP+CP	77.5%	80.3%	78.9%
BDT+SP+CP	**78.5%**	**81.5%**	**80.0%**

	Foreground	Background	Average
Bo & Fowlkes [21]	73.3%	81.1%	77.2%
Eslami & Williams [69]	71.6%	73.8%	72.7%
This work (BDT+SP+CP)	**78.5%**	**81.5%**	**80.0%**

Additional disparity gives another 2%.

Data:

- 300 images for superpixel classification training.
- ~10k pedestrian contours for MSSM.
- Penn-Fudan Test:
  169 pedestrian BBs.

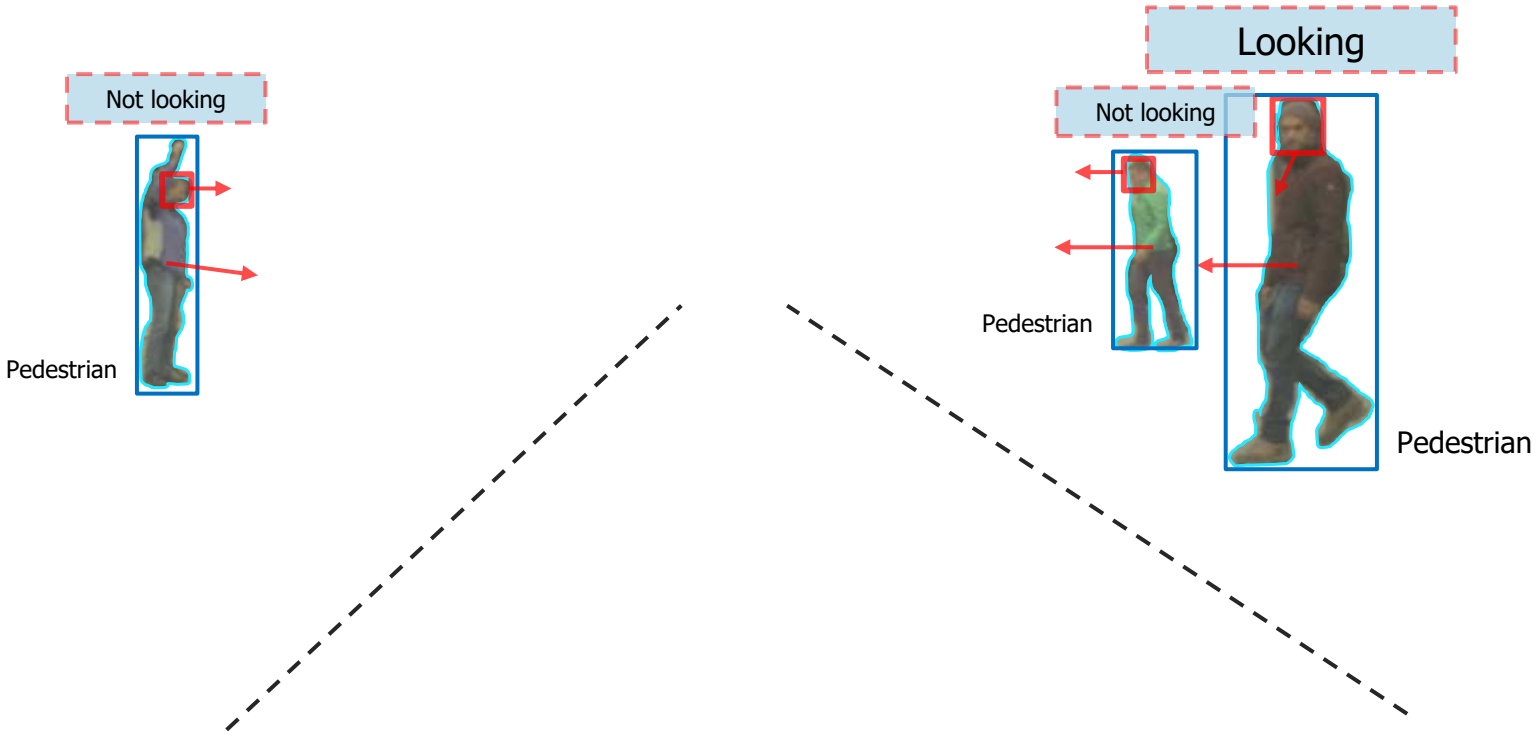
4

A challenge for vehicle automation

Orientation Estimation

Vulnerable Road User Orientation Estimation

Joint Head and Body Orientation Estimation



Empirical evidence suggests that the head and body orientation of a pedestrian are good indicators to a human what the pedestrian will do next [Hamaoka2013, Schmidt2009].

F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, and D. M. Gavrila, "Joint probabilistic pedestrian head and body orientation estimation," IV 2014

F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, and D. M. Gavrila, "A probabilistic framework for joint pedestrian head and body orientation estimation," T-ITS 2015

H. Hamaoka, T. Hagiwara, M. Tada, and K. Munehiro, "A study on the behavior of pedestrians when confirming approach of right/left-turning vehicle while crossing a crosswalk," IV 2013

S. Schmidt and B. Färber, "Pedestrians at the kerb - Recognising the action intentions of humans," Transportation Research, Part F 2009

Vulnerable Road User Orientation Estimation

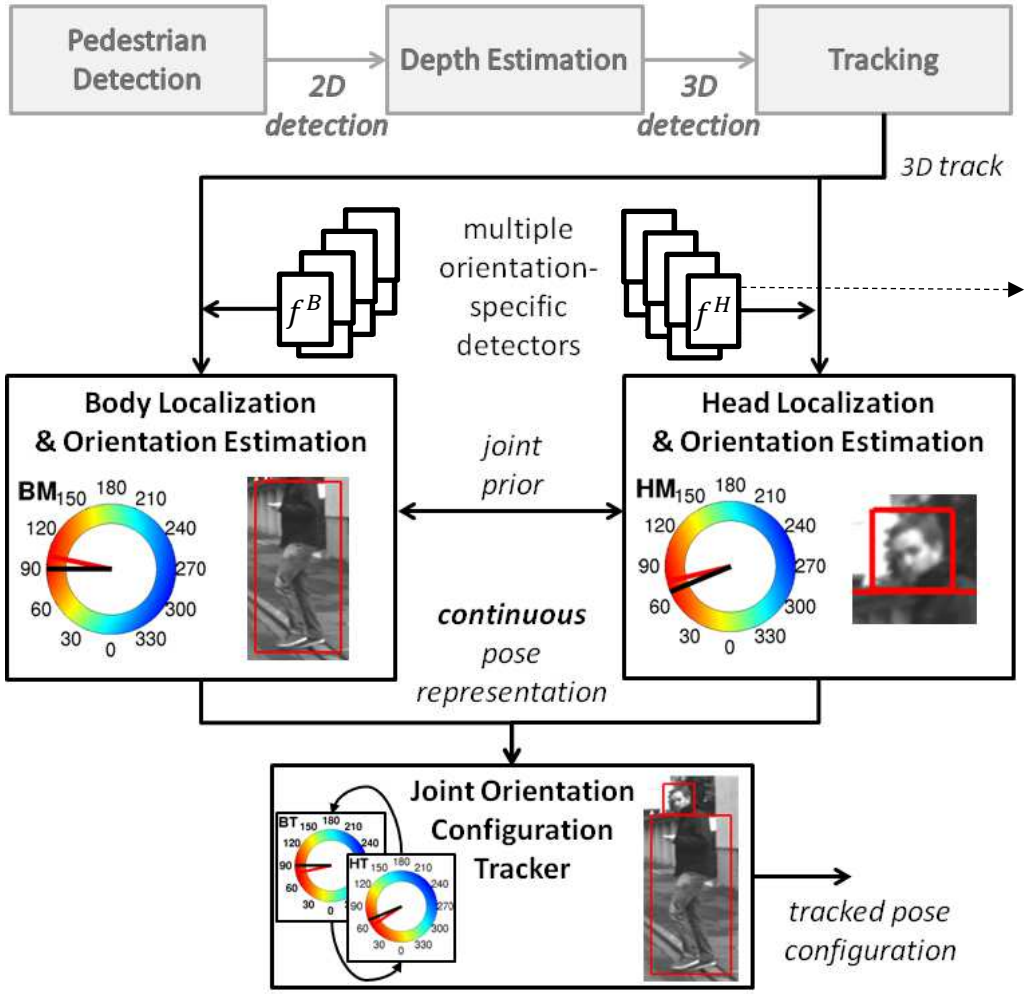
Contributions and Related Work

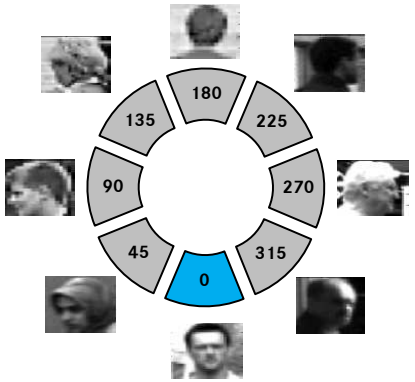
- **Joint** probabilistic **head** and **body orientation** estimation
  - handles **faulty part detections, continuous orientation** estimation and **temporal integration**
  - **constrained** coupling of body and head localization and orientation
- Experiments involve data from a vehicle-mounted stereo vision camera in a **realistic traffic setting**
  - Joint head an body orientation estimation reduces the mean absolute head and body orientation error **up to 11°** compared to single-frame estimation.
  - The system runs in **near-realtime** (8 – 9 fps)
- Differs to [Benfold2009, Schulz2011], where **no coupling** between head and body was used at all; [Rehder2014], where **no temporal or anatomical constraints** are used; [Benfold2009, Chen2012, Chen2011, Robertson2006], where **no intelligent vehicle context** is considered, introducing additional challenges like complex and ever-changing background and lighting conditions

B. Benfold and I. Reid, "Guiding visual surveillance by tracking human attention," BMVC 2009
A. Schulz, N. Damer, M. Fischer, and R. Stiefelhagen, "Combined head localization and head pose estimation for video-based advanced driver assistance systems", DAGM 2011
E. Rehder, H. Kloeden, and C. Stiller, "Head detection and orientation estimation for pedestrian safety", ITSC 2014
C. Chen and J. Odobez, "We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video", CVPR 2012
C. Chen, A. Heili, and J.-M. Odobez, "A joint estimation of head and body orientation cues in surveillance video," ICCV 2011
N. Robertson and I. Reid, "Estimating gaze direction from low-resolution faces in video", ECCV 2006

Vulnerable Road User Orientation Estimation

Joint Head and Body Orientation Estimation





Continuous orientation estimation of the parts (head, body) using set of orientation-specific detectors.

Coupling of body and head parts (image position and orientation).

Temporal integration (particle filtering, complete representation of the probability density function).

Vulnerable Road User Orientation Estimation

Filtering Theory

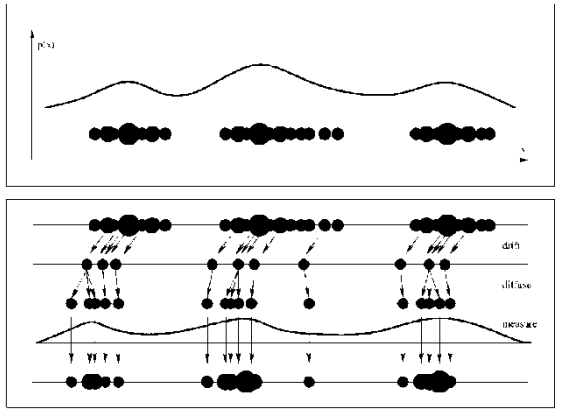
ω: orientation state (head, body)
z: orientation measurement
$\boldsymbol{x}$: position (velocity) measurement

$$p(\boldsymbol{\omega}_t|\boldsymbol{z}_{1:t}, \dot{\boldsymbol{x}}_{1:t}) \propto p(\boldsymbol{z}_t|\boldsymbol{\omega}_t)p(\boldsymbol{\omega}_t|\boldsymbol{z}_{1:t-1}, \dot{\boldsymbol{x}}_{1:t})$$

State estimate at t | Observation Model | State Prediction

$$p(\boldsymbol{\omega}_t|\boldsymbol{z}_{1:t-1}, \dot{\boldsymbol{x}}_{1:t}) = \int p(\boldsymbol{\omega}_t|\boldsymbol{\omega}_{t-1}, \dot{\boldsymbol{x}}_t)p(\boldsymbol{\omega}_{t-1}|\boldsymbol{z}_{1:t-1}, \dot{\boldsymbol{x}}_{1:t-1})d\boldsymbol{\omega}_{t-1}$$

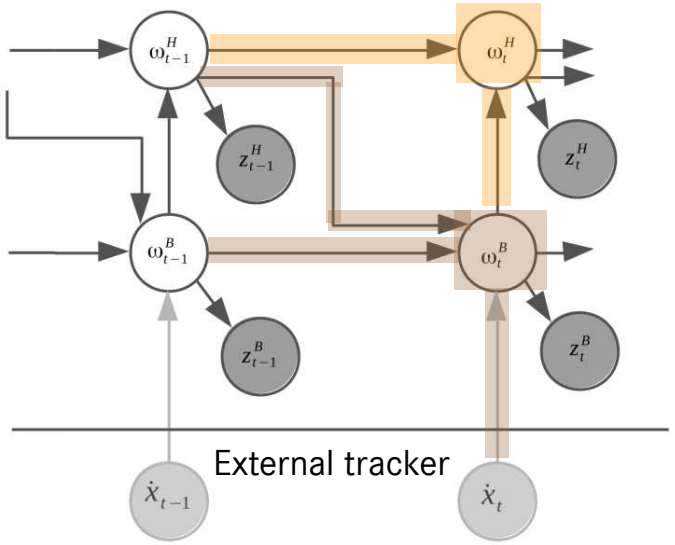
Dynamical Model | State estimate at t-1



Multimodality → inference by particle filtering

Vulnerable Road User Orientation Estimation

Dynamic model



Conditional independence of head and body orientation

$$p(\boldsymbol{\omega}_t|\boldsymbol{\omega}_{t-1}, \dot{\boldsymbol{x}}_t) = p(\omega_t^H|\omega_{t-1}^H, \omega_t^B)p(\omega_t^B|\omega_{t-1}^B, \omega_{t-1}^H, \dot{\boldsymbol{x}}_t).$$

Head orientation: temporal constraints and coupling to body orientation

$$p(\omega_t^H|\omega_{t-1}^H, \omega_t^B) = \alpha_{hh}\mathcal{V}(\omega_t^H; \omega_{t-1}^H, \kappa_{hh}) + (1-\alpha_{hh})\mathcal{V}(\omega_t^H; \omega_t^B, \kappa_{hb})$$

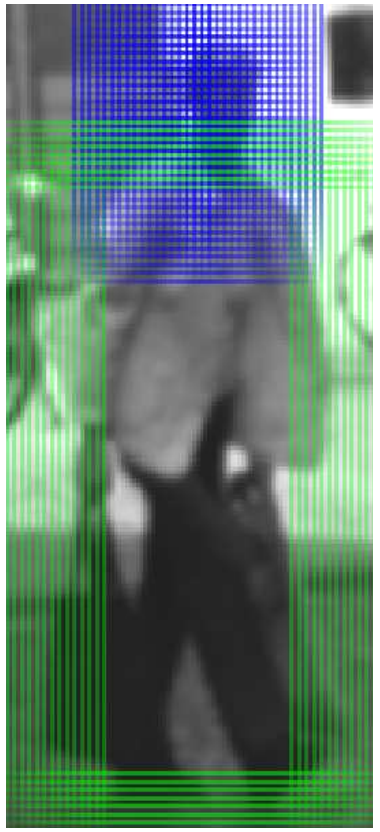
Body orientation: temporal constraints, coupling to body orientation, and dependence on movement direction.

$$p(\omega_t^B|\omega_{t-1}^B, \omega_{t-1}^H, \dot{\boldsymbol{x}}_t) = \alpha_{bb}\mathcal{V}(\omega_t^B; \omega_{t-1}^B, \kappa_{bb}) + \alpha_{bh}\mathcal{V}(\omega_t^B; \omega_{t-1}^H, \kappa_{bh}) + (1-\alpha_{bb}-\alpha_{bh})\mathcal{V}(\omega_t^B; ang(\dot{\boldsymbol{x}}_t), \kappa_{bv})$$

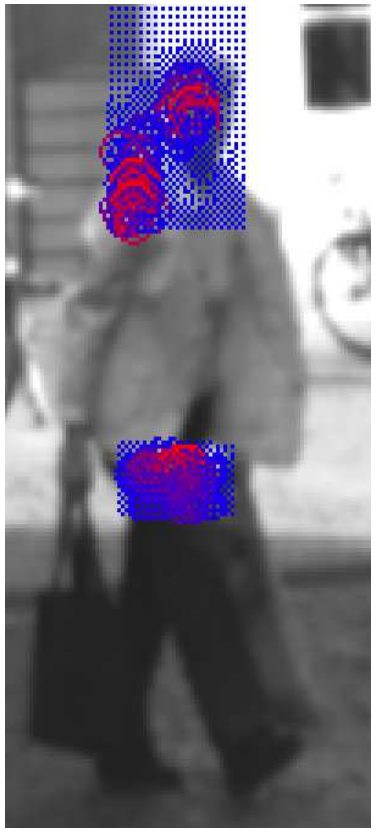
The von Mises distribution $\mathcal{V}(\cdot;\omega,\kappa)$ is the analogue of the normal distribution for the circular domain, with mean angle ω and concentration parameter κ („sigma“) . It reduces to a circular uniform distribution for $\kappa = 0$.

Vulnerable Road User Orientation Estimation

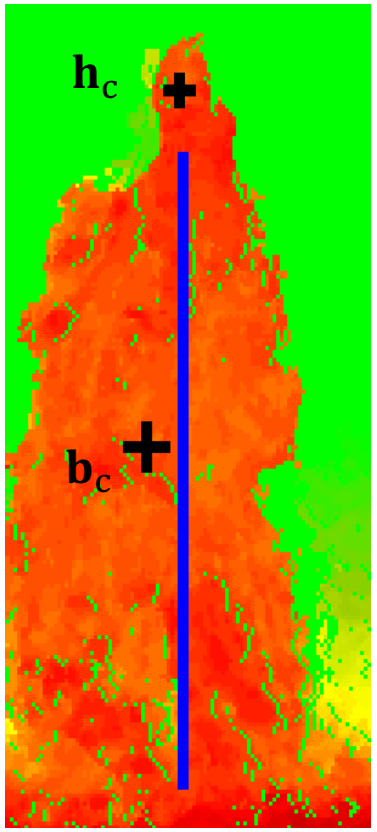
Observation model



Region candidates



Region probabilities



Stereo likelihoods

Structure knowledge

Vulnerable Road User Orientation Estimation

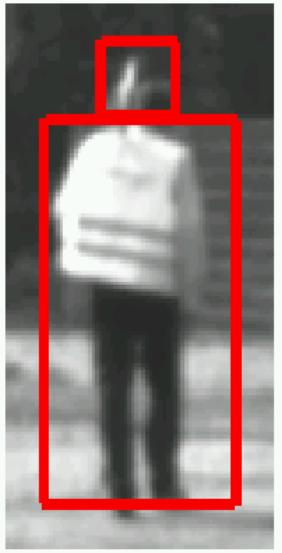
Help of Anatomical Constraints and Temporal Filtering

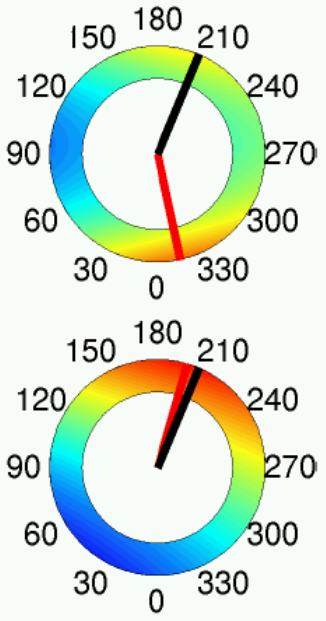
Joint body and head localization

independently

joint
(spatial constraints)

Temporal filtering





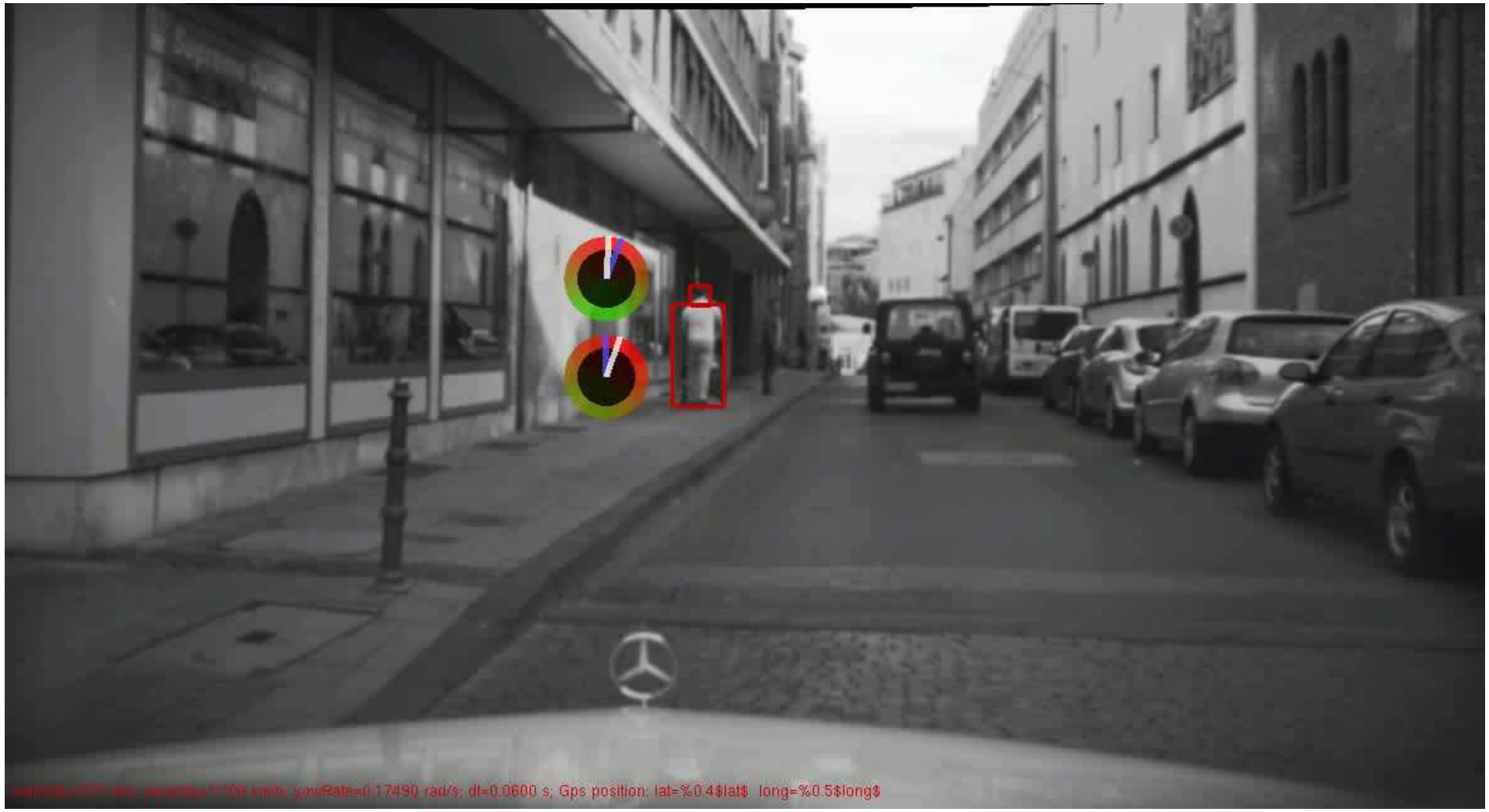
Based on measurement only (wrong ML)

Final estimate (correct MAP)

Red line: maximum likelihood/posterior estimate
Black line: ground truth orientation
Heat map coloring encodes probability density.

Vulnerable Road User Orientation Estimation

Qualitative Results

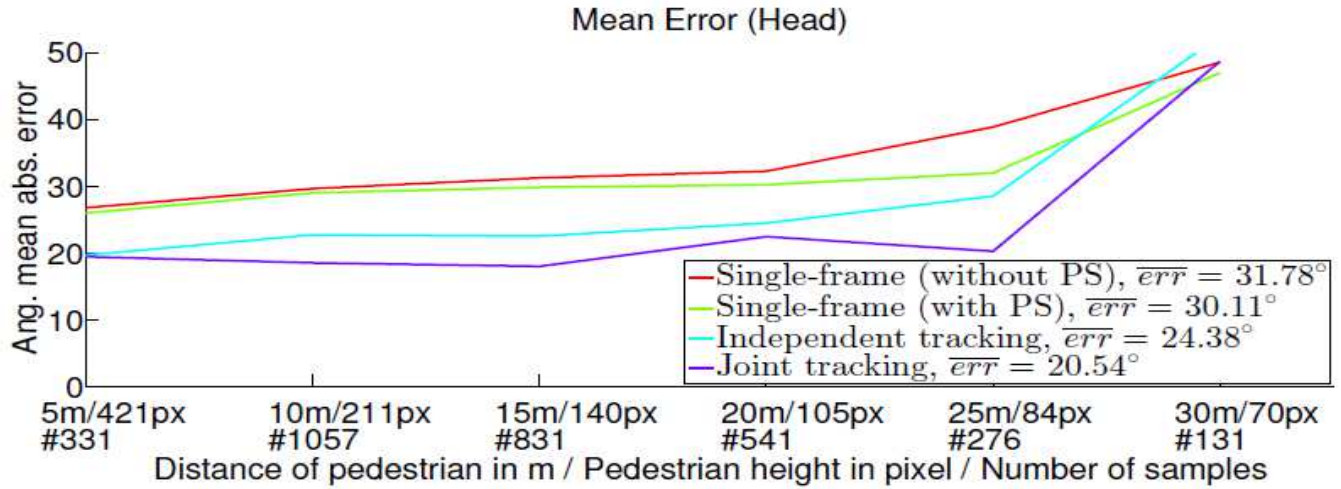


Head and torso estimation takes into account physical constraints.

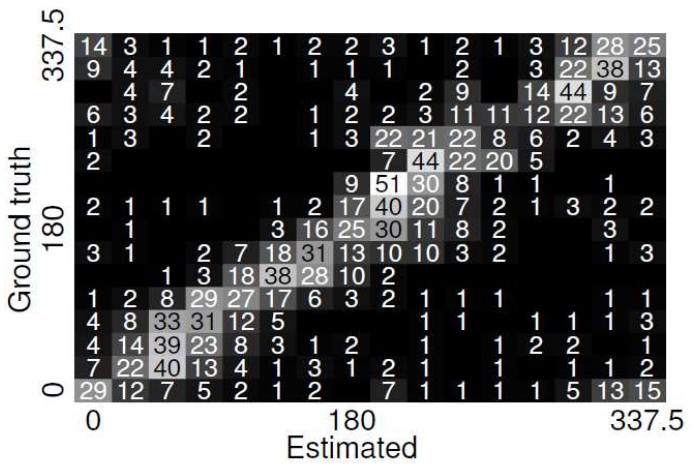
For clarity, results are only shown for one pedestrian at a time.

Vulnerable Road User Orientation Estimation

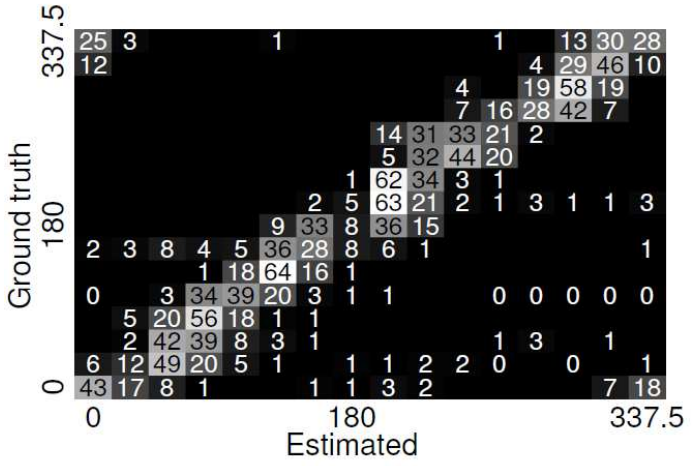
Quantitative Results



(a) Angular mean absolute error for head (in degrees)



(a) Head (single-frame)



(c) Head (tracked)

Data:

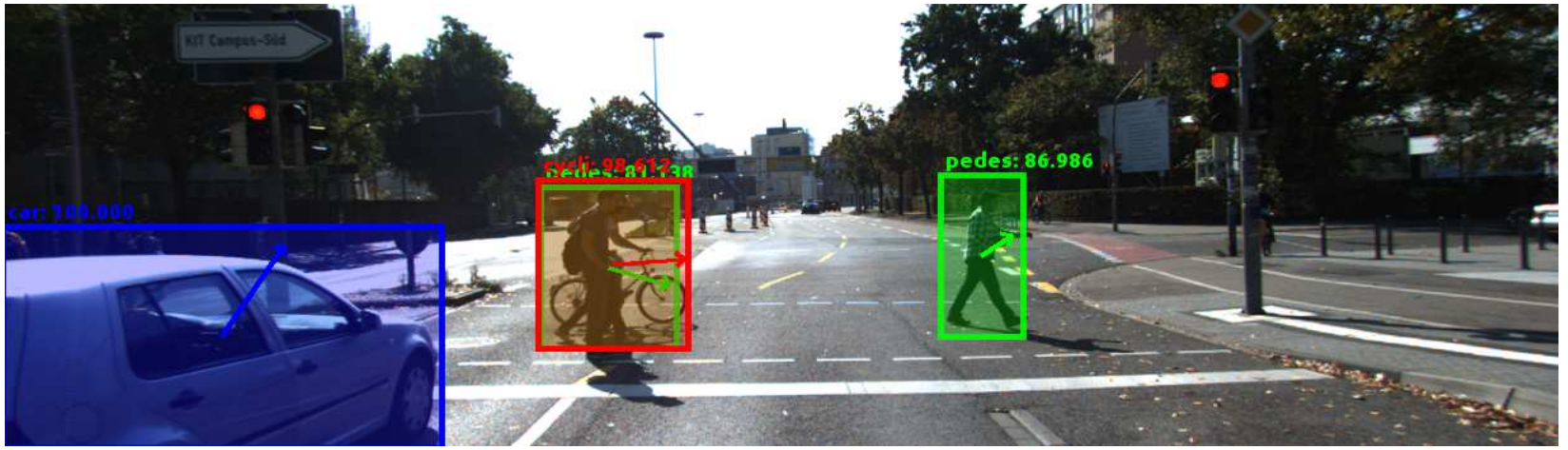
- 9300 head/body cut-outs for classifier training
- 32/60 real-world sequences for validation / test

Robust results up to 25m

Error reduction of about 11° compared to single-frame estimation

Vulnerable Road User Joint Detection and Orientation Estimation

Using Von Mises loss in a Fast-RCNN architecture



Robust object orientation is estimated together with bounding box classification and regression.

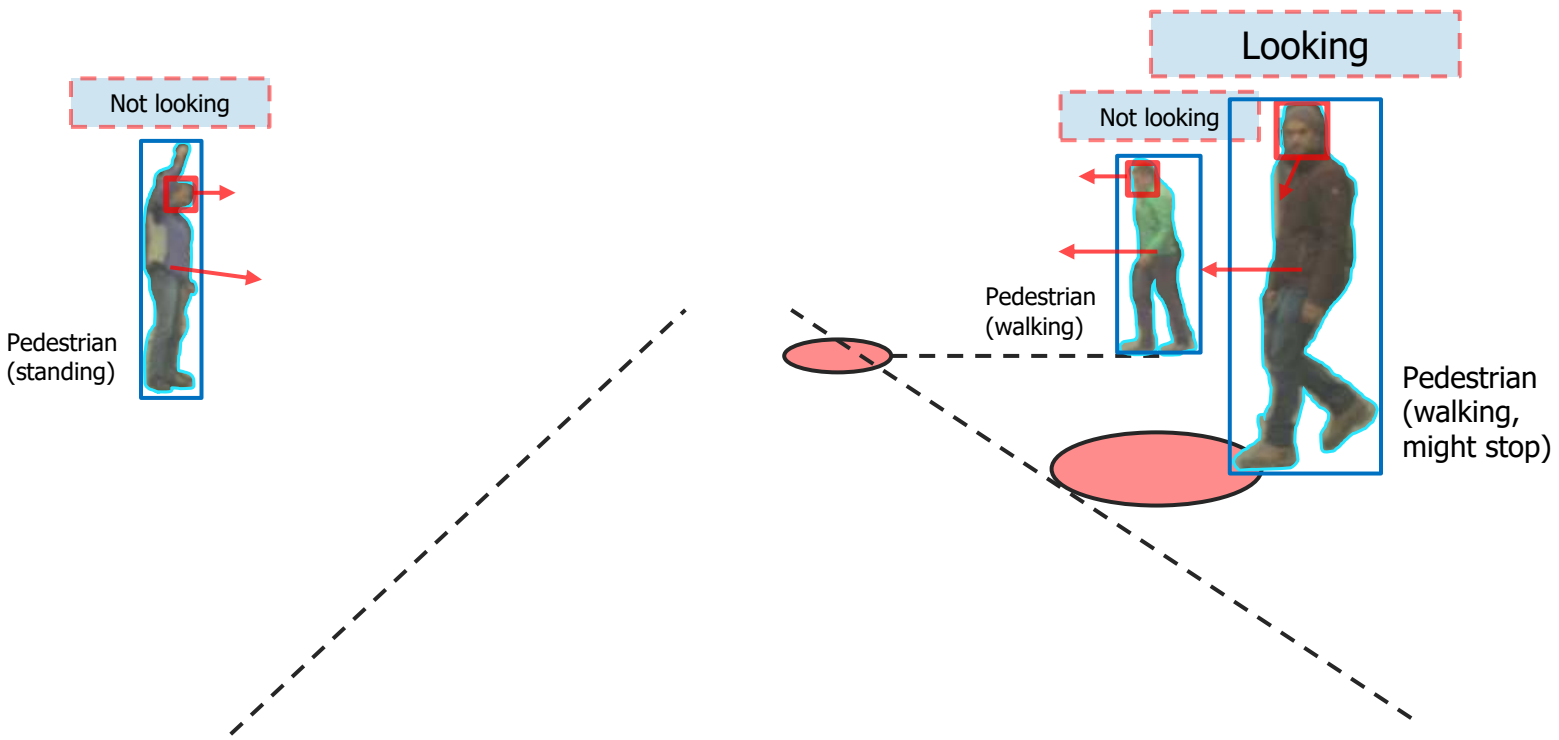
M. Braun, Q. Rao, Y. Wang and F. Flohr. *"Pose-RCNN: Joint object detection and pose estimation using 3D object proposals"*, *ITSC* 2016

Where vehicle automation gets complicated

Predicting the VRU

Vulnerable Road User Intention Recognition and Path Prediction

Context-based pedestrian path prediction



Extracted visual VRU context cues (e.g. by means of the head orientation) are used to improve VRU path prediction.

J. P. F. Kooij, N. Schneider, F. Flohr and D. M. Gavrila. *"Context-based pedestrian path prediction"*. ECCV 2014

J. F. P. Kooij, F. Flohr, E.A.I. Pool and D.M. Gavrila. *"Context-based Path Prediction for Targets with Switching Dynamics"*, IJCV 2018

Vulnerable Road User Intention Recognition and Path Prediction

Contributions and Related Work

- Realtime system integration using a **Context-based Switching Linear Dynamic System** (C-SLDS) formulation [Kooij2014] to enhance path prediction.
- **Head orientation** is used to model **pedestrian's awareness** [Flohr2015]
- Evaluations show the benefit of VRU behavior cues integrated in the C-SLDS approach.
  - The prediction horizon can be increased **up to 1 s** without increasing the false alarm rate.
  - Compared to other methods [e.g. Keller2014, Fang2017] the integrated systems runs in **realtime** with 16fps
  - In the **online demonstration** the system is compared to a simpler Linear Dynamic System (LDS) [Blackman1999] to show the benefit for an improved acoustical driver warning.

J. P. F. Kooij, N. Schneider, F. Flohr and D. M. Gavrila. *"Context-based pedestrian path prediction"*. ECCV 2014

F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, and D. M. Gavrila, "A probabilistic framework for joint pedestrian head and body orientation estimation," T-ITS 2015

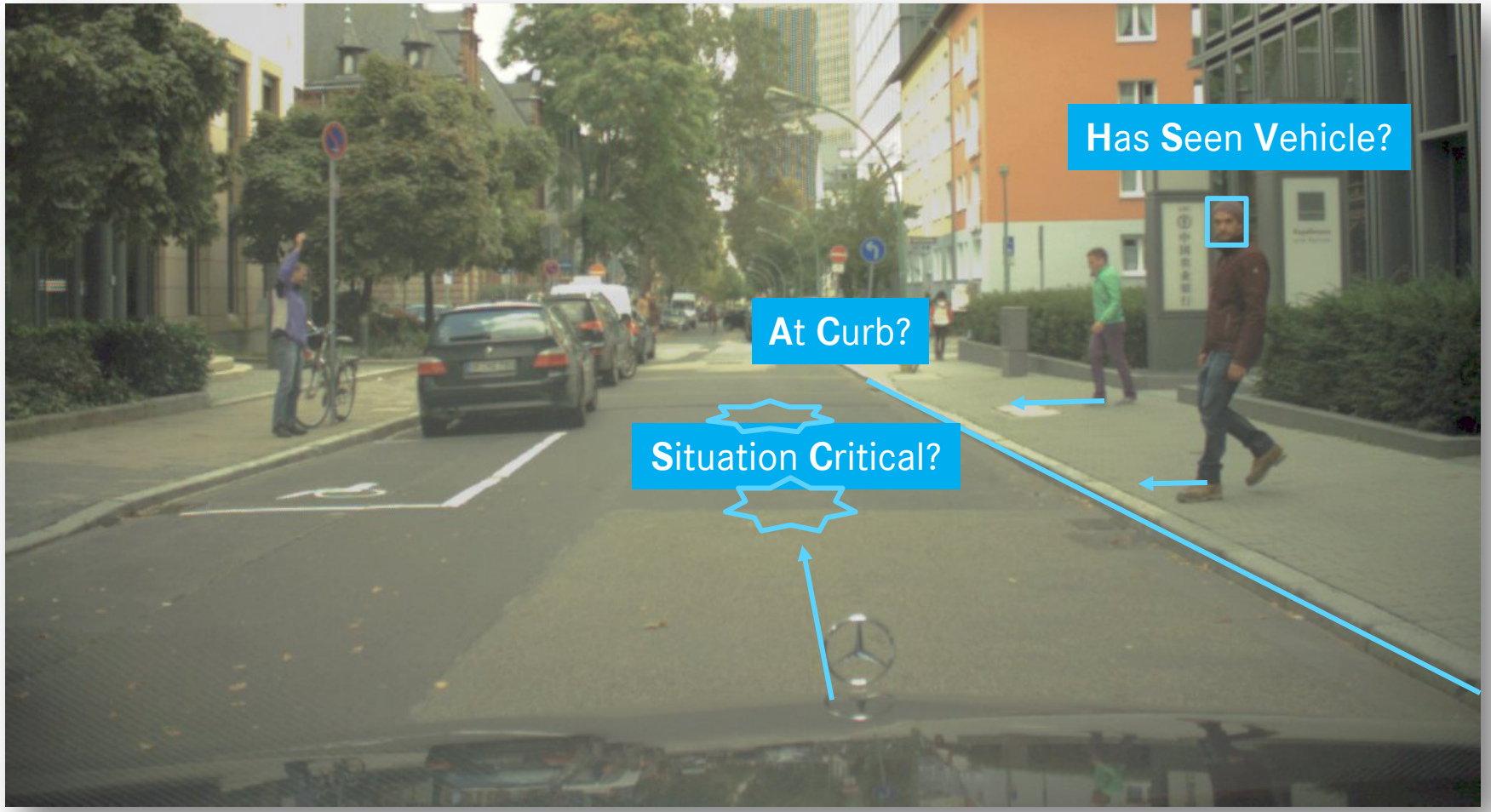
S. Blackman and R. Popoli, *"Design and analysis of modern tracking systems"*, Artech House Norwood, 1999

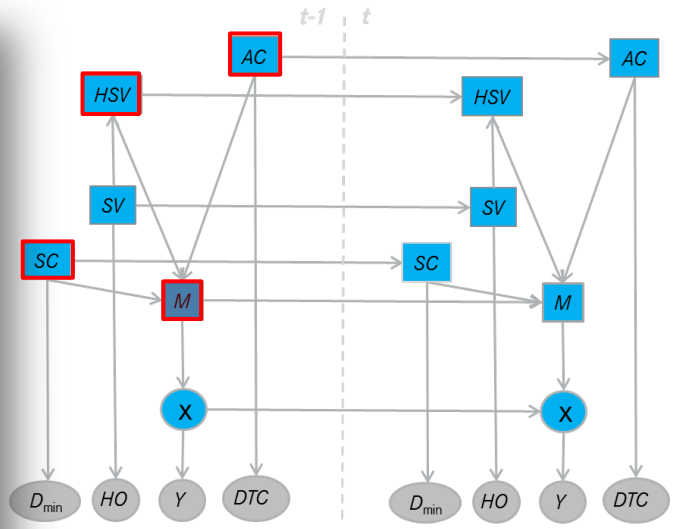
C. G. Keller and D. M. Gavrila, *"Will the pedestrian cross? A study on pedestrian path prediction,"* T-ITS 2014

Z. Fang, D. Vázquez, and A. M. López, *"On-board detection of pedestrian intentions,"* Sensors (Basel, Switzerland) 2017

Vulnerable Road User Intention Recognition and Path Prediction

Dynamic Bayesian Networks with Context Cues





Pedestrian Motion Modeling: Dynamic Bayesian Networks.

Various context cues influence the pedestrian motion.

Vulnerable Road User Intention Recognition and Path Prediction

Dynamic Bayesian Networks with Context Cues

A human driver relies heavily on context cues to anticipate how a traffic situation will evolve.

Various context cues influence the pedestrian motion.

Vulnerable Road User Intention Recognition and Path Prediction

„Live“ demo (critical crossing case)

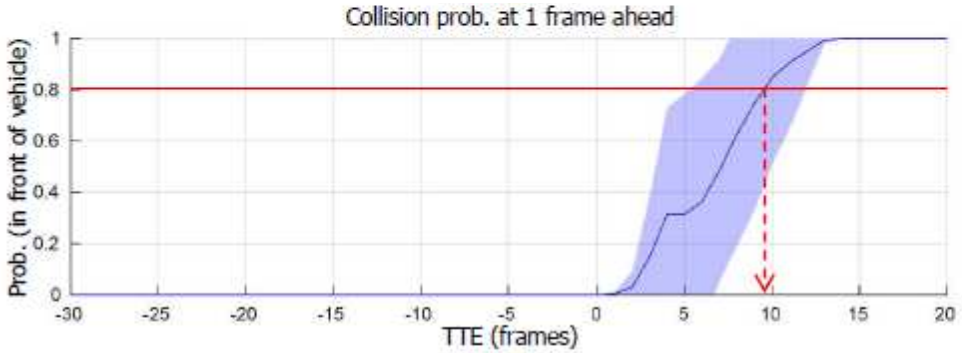
High tone: „production system“ warning, based on pedestrian point kinematics

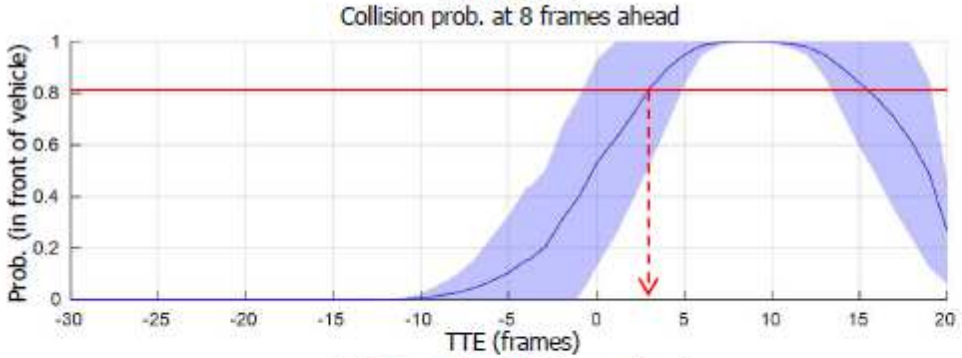
Low tone: „new“ warning, based on context-based pedestrian motion modeling

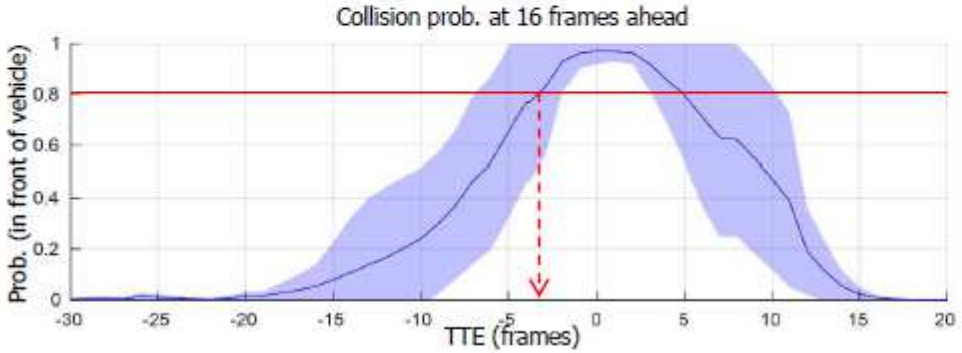
„New“ warning comes 1 second earlier !

Vulnerable Road User Intention Recognition and Path Prediction

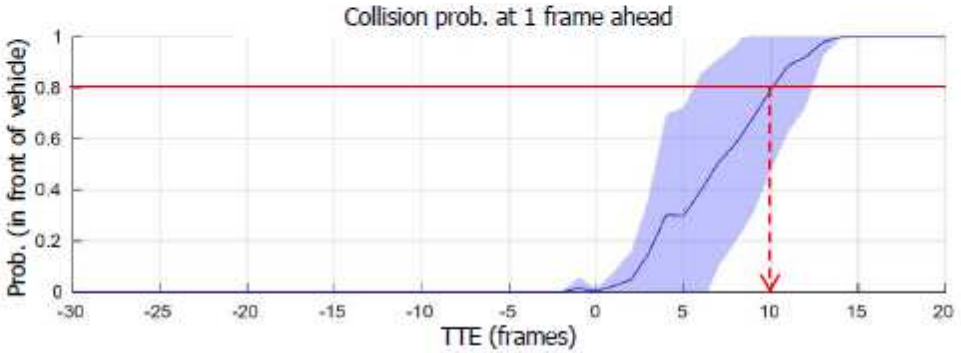
Evaluation of the critical crossing case

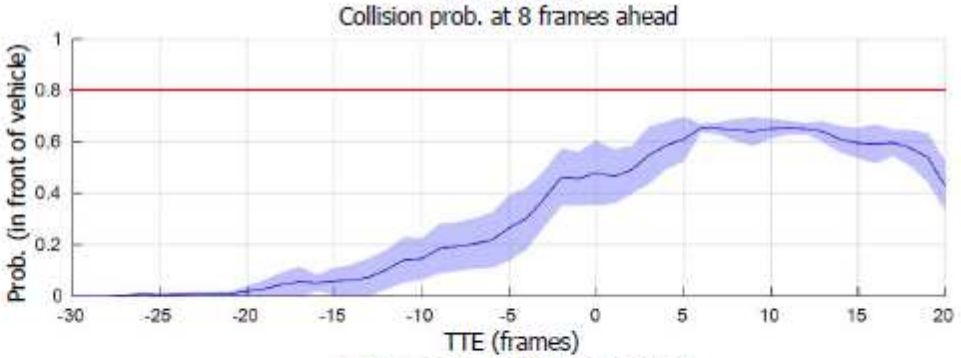


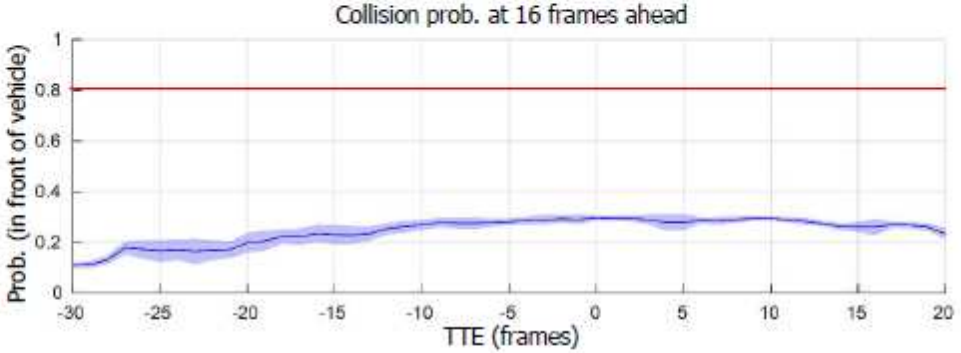




(a) C-SLDS







(b) LDS

C-SLDS is able to warn robustly about 16 frames (~1s) earlier.

Uncertainty in a LDS is too high when using greater prediction horizons.

Data:

- 48 “normal” sequences *Leave-one-out cross-validation*

Vulnerable Road User Intention Recognition and Path Prediction

„Live“ demo (critical stopping case)

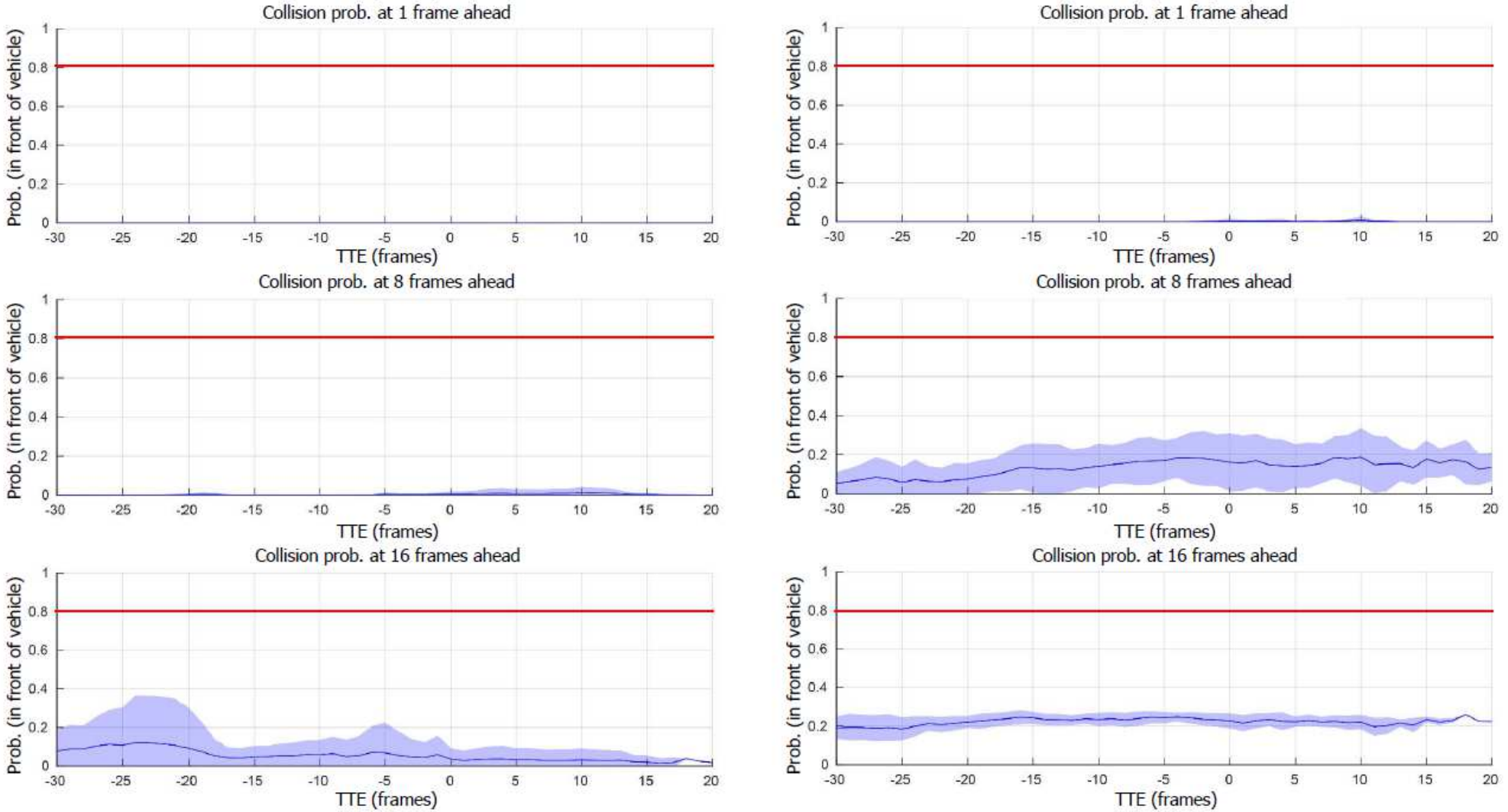
High tone: „production system“ warning, based on pedestrian point kinematics

Low tone: „new“ warning, based on context-based pedestrian motion modeling

No false alarm !

Vulnerable Road User Intention Recognition and Path Prediction

Evaluation of the critical stopping case



(a) C-SLDS (b) LDS

C-SLDS correctly recognizes that the pedestrian will stop.

Uncertainty in a LDS is too high when using greater prediction horizons.

6

How this leads finally to an

Understanding the VRU

(conclusion / future work)

Vulnerable Road User Understanding

Conclusions and Lessons Learned

- VRU detection performance can benefit from an **efficient stixel-based proposal** method.
- Combining **top-down shape** information with **bottom-up data cues** helps to cope with large variations in VRU appearance across cluttered background.
- Efficient part localization and orientation estimation can be done by applying a set of **domain-specific detectors** guided by **anatomical and temporal constraints**.
- The use of **visual context cues** allows the application of multiple linear motion models to gain a more **accurate path prediction** of the VRU.
- **Concrete system integrations** and demonstrations help to evaluate research work and provide valuable feedback for improvement and the direction of future work.
- **Efficient data labeling and management** for developing machine learning applications are often underrated tasks and deserve more attention in research.
- **Understanding the VRU** is an enabler for a seamless integration of autonomous cars into our daily environment…

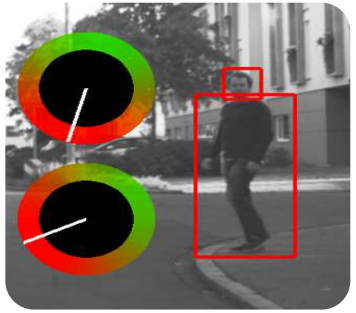
Vulnerable Road User Understanding

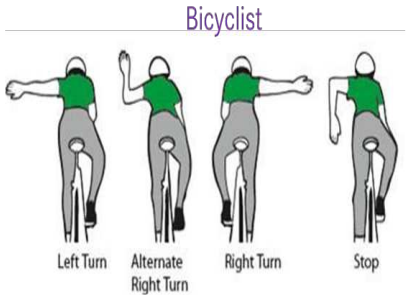
Why Do We Need to Understand Gestures?

Needed for

Anticipation of VRU Behavior

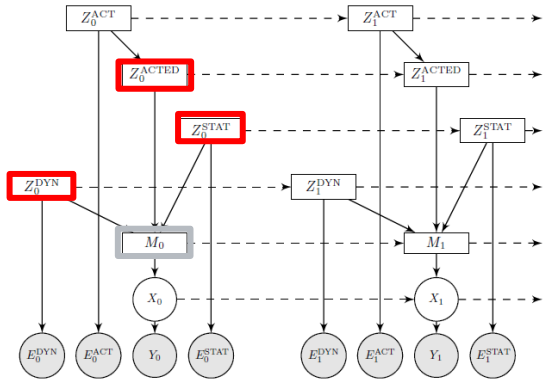
- Gestures are used as VRU behavior cues
- VRU behavior cues
  - ➢ indicate changes in VRU motion
  - ➢ allow a higher prediction horizon of VRUs





Vulnerable Road User Understanding

Application of cyclist gestures for path prediction



At intersection?
On collision course?
Hand gesture?

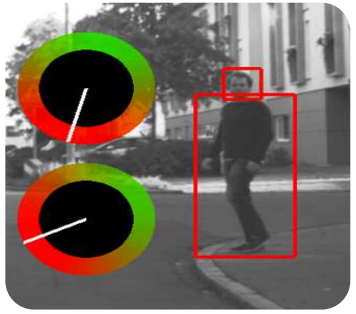
J.F.P. Kooij, F. Flohr, E.A.I. Pool and D.M. Gavrila. *"Context-based Path Prediction for Targets with Switching Dynamics"*, IJCV 2018

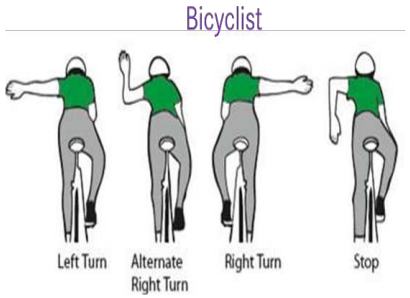
Vulnerable Road User Understanding

Why Do We Need to Understand Gestures?

Needed for
Anticipation of VRU Behavior

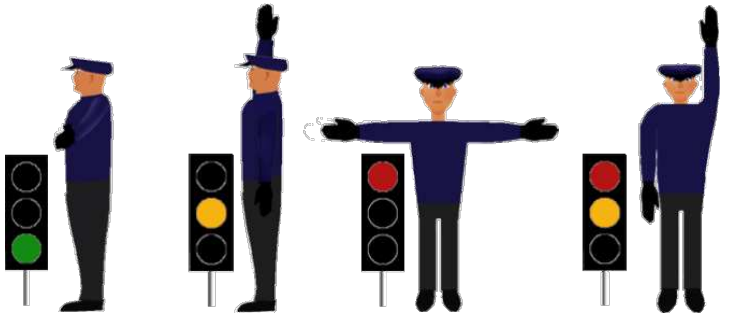
- Gestures are used as VRU behavior cues
- VRU behavior cues
  - ➢ indicate changes in VRU motion
  - ➢ allow a higher prediction horizon of VRUs





Needed for
Understanding Traffic Rules or for Interaction with AV

- Special persons are allowed to change or replace traffic rules
- Interaction with AV becomes necessary (e.g. calling taxi)



Vulnerable Road User Understanding

How Can We Understand Gestures?

Real-time Multi-Person 2D Pose Estimation Using Part Affinity Fields

Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh

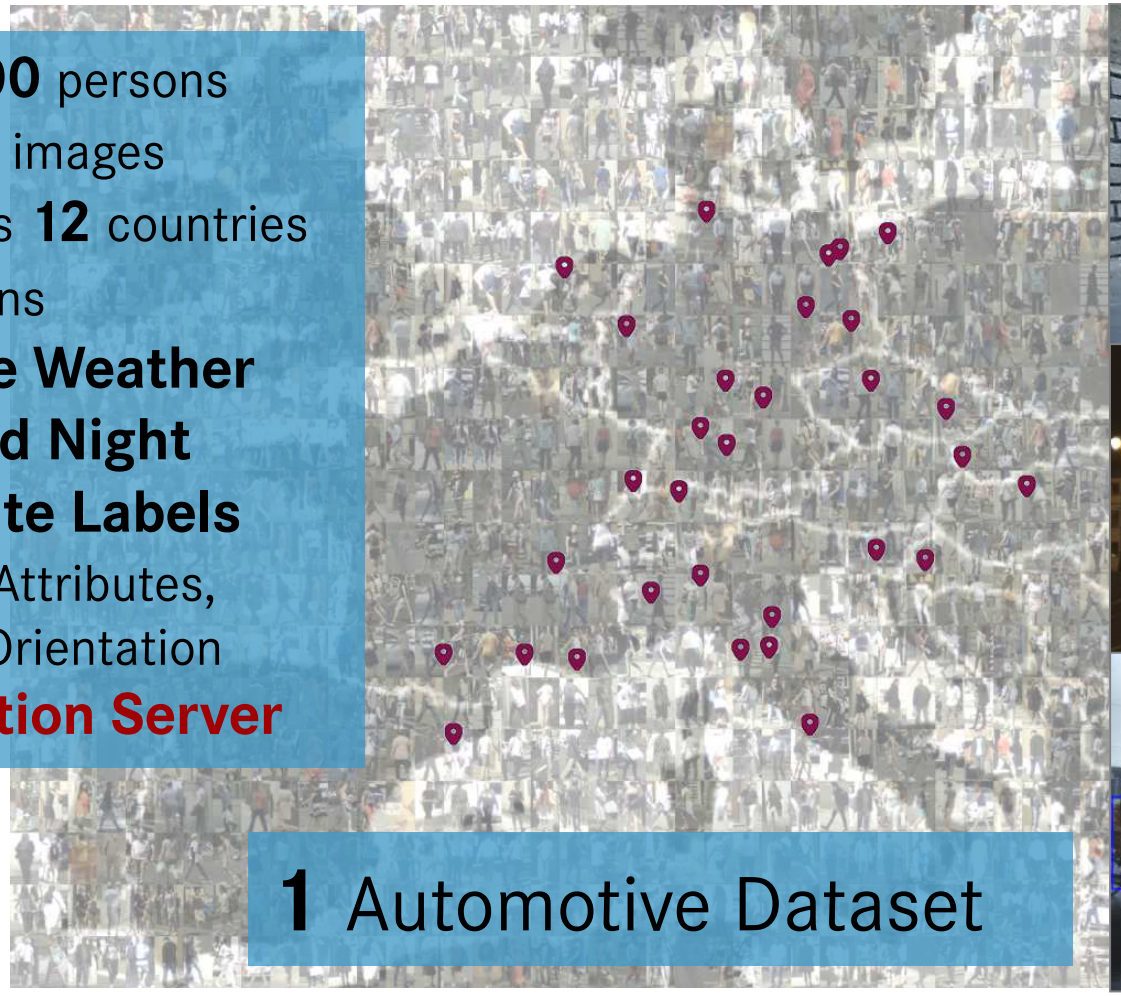
Carnegie Mellon University

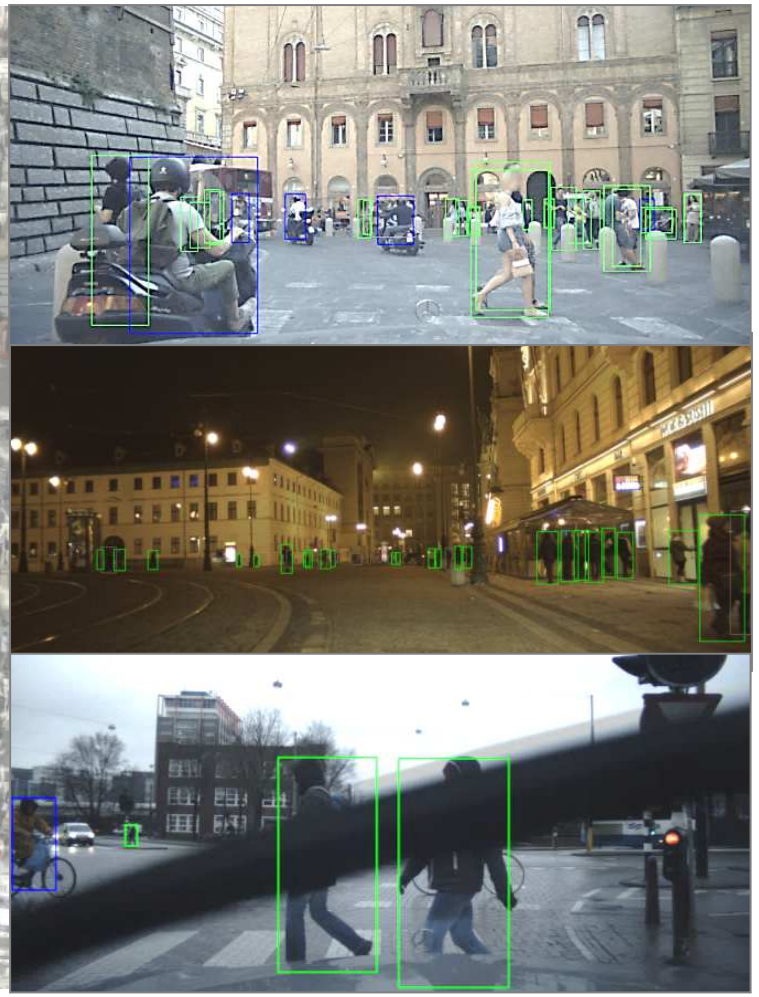
Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, *"Realtime multi-person 2d pose estimation using part affinity fields,"* CVPR 2017

Feeding the Brain More Data

The EuroCity Person Detection Dataset

238.200 persons
47.300 images
31 cities **12** countries
4 seasons
Diverse Weather
Day and Night
Accurate Labels
Class, Attributes,
Bbox, Orientation
Evaluation Server





1 Automotive Dataset

M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila *"The EuroCity Persons Dataset: A novel benchmark for object detection"*, PAMI 2018 [in revision]

Vehicle Automation Has a Deeper Impact

Vehicle automation opens up new vehicle concepts.

New Interior/UI Concepts

New Signalling Concepts

Thank You Very Much For Your Attention!