Scene Perception based on Boosting over Multimodal Channel Features

Arthur Costea

Image Processing and Pattern Recognition Research Center

Technical University of Cluj-Napoca

Research Group

Technical University of Cluj-Napoca, Romania

Image Processing and Pattern Recognition Research Center

http://cv.utcluj.ro/

Coordinator: Prof. Dr. Eng. Sergiu Nedevschi

Assoc. Prof. Dr. Eng. Tiberiu Mariţa
Assoc. Prof. Dr. Eng. Radu Dănescu
Assoc. Prof. Dr. Eng. Florin Oniga
Assist. Prof. Dr. Eng. Delia Mitrea
Assist. Prof. Dr. Eng. Cristian Vicas

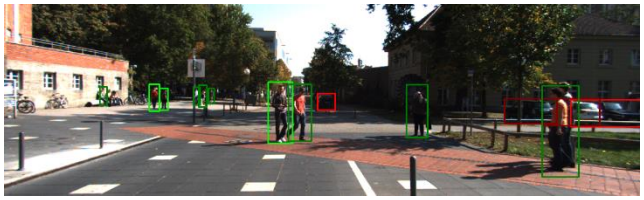
Assist. Dr. Inf. Anca Ciurte
Assist. Dr. Eng. Andrei Vatavu
Assist. Dr. Eng. Ion Giosan
Assist. Dr. Eng. Raluca Brehar
Assist. Dr. Eng. Mihai Negru
Assist. Dr. Eng. Ciprian Pocol
Dr. Eng. Pangyu Jeong

PhD Student Catalin Golban
PhD Student Cristian Vancea
PhD Student Marius Drulea
PhD Student Robert Varga
PhD Student Vlad Miclea
PhD Student Andra Petrovai
PhD Student Mircea Muresan
PhD Student Claudiu Decean
PhD Student Arthur Costea

Overview

- Perception tasks:
  - Object detection
  - Semantic segmentation

- Objectives
  - High recognition accuracy and precision
  - Fast execution time
  - Enable real-time detection on mobile devices



Overview

- Common framework for detection and segmentation:
  - Features: image channels
    - Word Channels
    - Multiresolution Filtered Channels
    - Semantic Channels
    - Multimodal Channels
    - Deep Convolutional Channels
  - Classification: boosting over channel features
    - Easy fusion of different features types
    - Low computational costs

EU Research Projects

- **CoMoSeF** – Co-operative Mobility Services of the Future

  Celtic Plus EU project (2012-2015)

- **PAN-Robots** – Plug & Navigate robots for smart factories

  FP7 EU project (2012-2015)

PAN ROBOTS
Plug&Navigate robots for smart factories

- **UP-Drive** – Automated Urban Parking and Driving

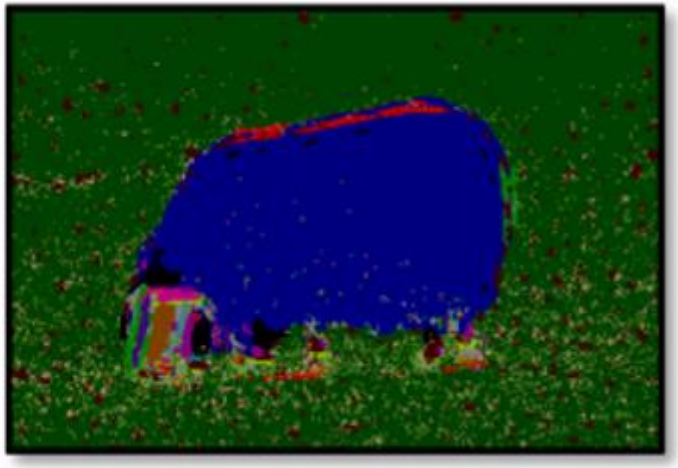
  H2020 EU project (2016-2019)

UP-Drive

Word Channels

- Visual codebook based image representation
- Image is represented as a distribution of visual words

Input



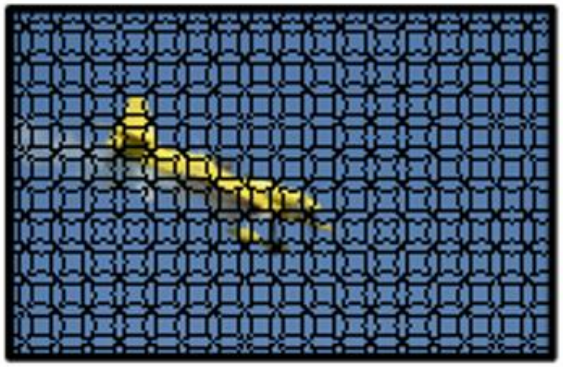
Texton Map

[Shotton et al. 2006]

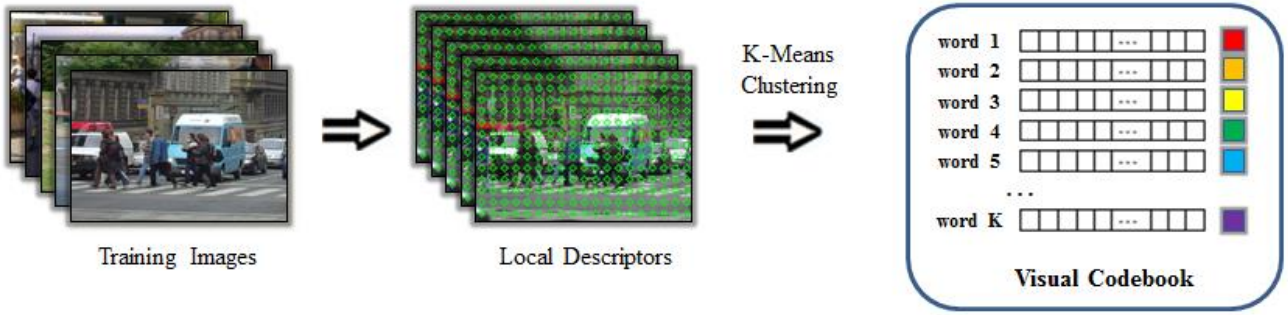
Word Channels

Local Descriptors:

- Describe a local neighborhood of pixels
- We employ three descriptor types:
  - HOG, LBP and color
- Dense sampling of descriptors (pixelwise)

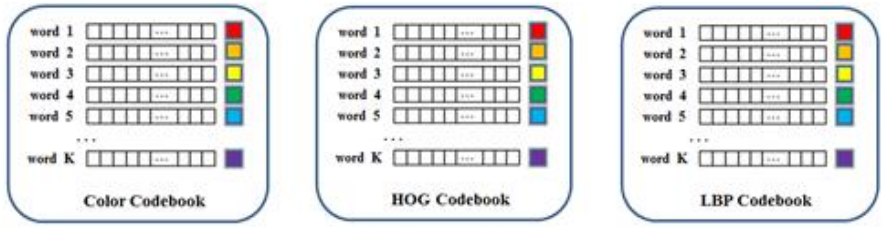


Visual Codebooks: a collection of descriptor vectors



Word Channels

Codebook mapping:



Input Image

Local descriptors for each descriptor type

Mapping of local descriptors to codebook words

Word Channels:

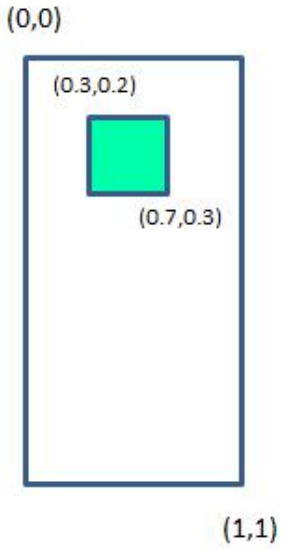
Color

HOG

LBP

Pedestrian classification

- Shape filter:
  - One codebook word
  - Rectangle (relative position and size)
- Shape filter response:
  - Normalized codebook word count inside the rectangle



Pedestrian classification

- Detection window classification:
  - Pedestrian vs. Non-pedestrian
- Classification features:
  - Shape filter responses
  - $|\boldsymbol{S}|$ x $|\boldsymbol{F}|$ features
- Classifier:
  - Boosted decision stumps over shape filter responses
  - 1000 boosting rounds
- Train a cascade of boosting classifiers

Multiscale detection

- Multiscale sliding window based detection

1 model, N image scales
Traditional approach

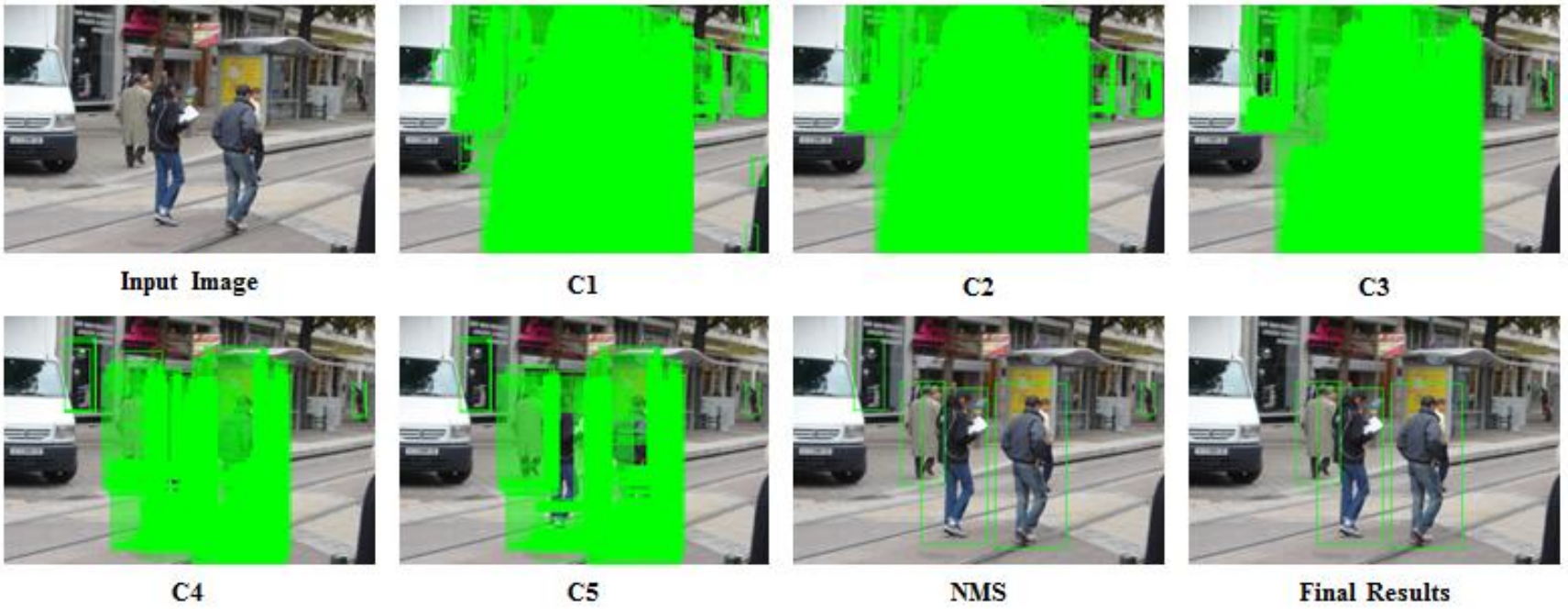
1 model, N/K image scales
FPDW approach

N/K models, 1 image scale
VeryFast approach

1 model, 1 image scale
Our approach

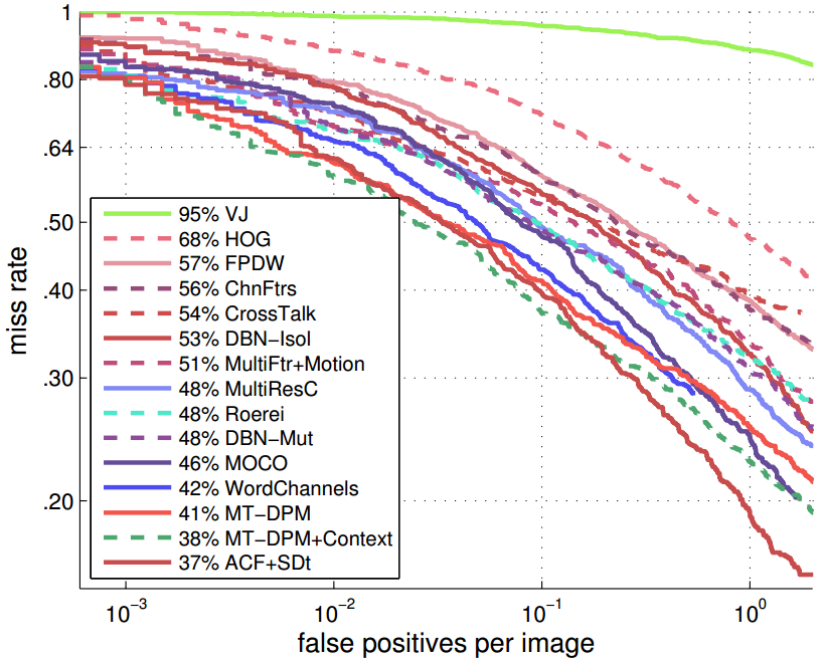
Pedestrian detection

Cascade classification:

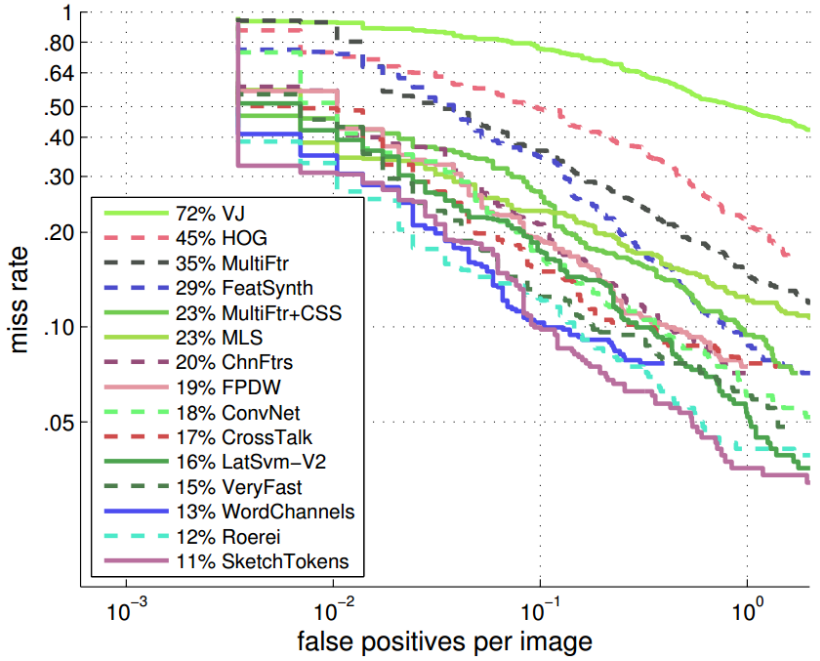


Pedestrian detection evaluation

Caltech – reasonable



INRIA



(2014)

Computational costs

- Average execution times for 640 x 480 images:
  (GPU implementation on an Nvidia 780 GTX)
  - Pixel-wise local descriptor computation: 4 ms
  - Codebook matching: 8 ms
  - Integral image computation: 11 ms
  - Classification of each bounding box: 39 ms
- Total detection time: 62 ms (16 FPS)
- Total training time: ~30 minutes

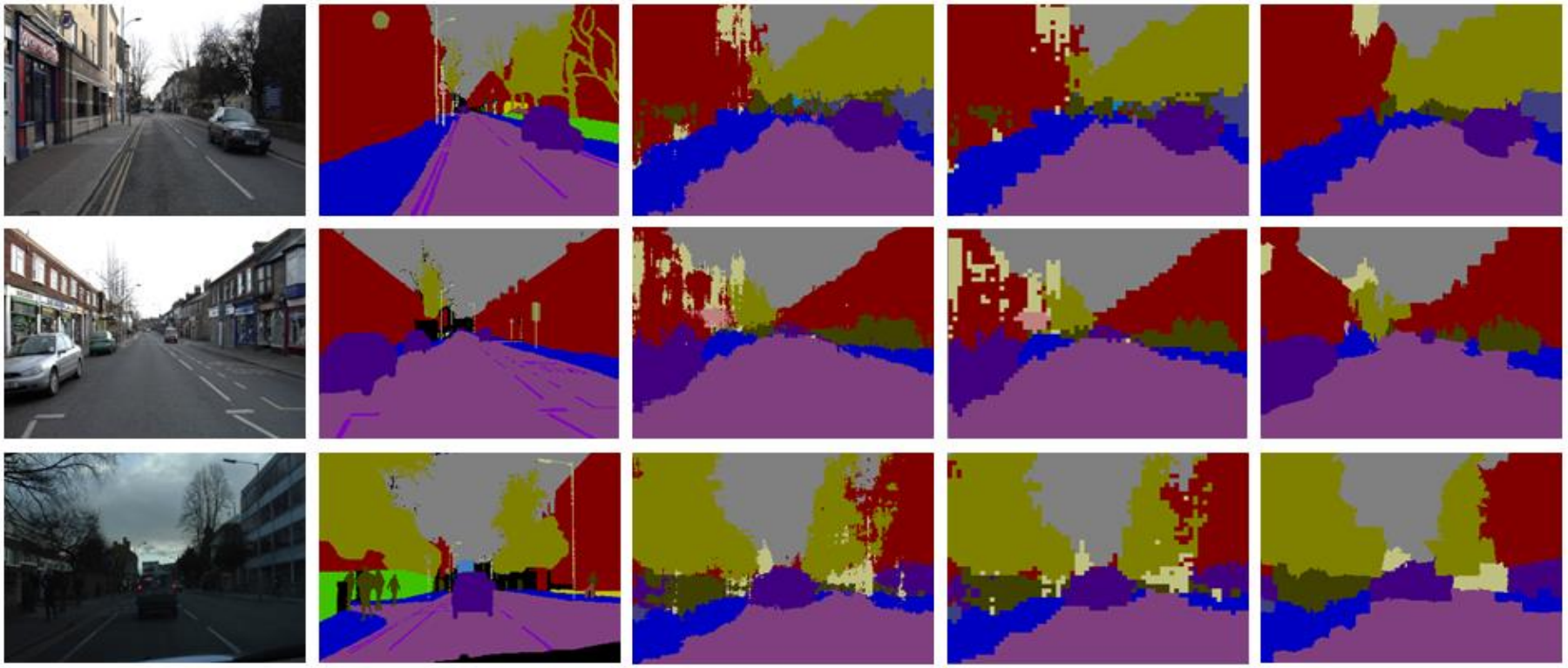
Pixel classification

Word Channel feature based pixel classification:

- Similar classification scheme
- A pixel is classified based on surrounding visual words
- Use of 100 random rectangles inside of a 200x200 pixel region for learning (*TextonBoost* [Shotton et al. 2006])
- Classifier:
  - Multi-class boosted decision stumps => joint boosting
  - 4096 boosting rounds

Multi-class segmentation results

CamVid segmentation benchmark



Input Image | Ground Truth | Pixel classification SS1 | Pixel classification SS5 | Superpixel CRF Unary + Pairwise

Segmentation evaluation

CamVid segmentation benchmark:

	FPS	**Global**	**Average**	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Sidewalk	Bicyclist
Brostow et al. (Motion) [4]	1	61	43	43	46	79	44	19	82	24	58	0	61	18
Brostow et al. (Appearance) [4]		66	52	38	60	90	71	51	88	54	40	1	55	23
Brostow et al. (Combined) [4]		69	53	46	61	89	68	42	89	53	46	0	60	22
Our - Unary pixel – SS1	14	74	53	60	77	82	72	8	92	53	27	29	62	19
Our - Unary pixel – SS5	65	72	53	52	73	82	73	7	90	62	29	31	67	17
Our - Unary superpixel (SS5) + Smoothness	36	76	52	66	81	84	71	2	94	50	25	20	60	13

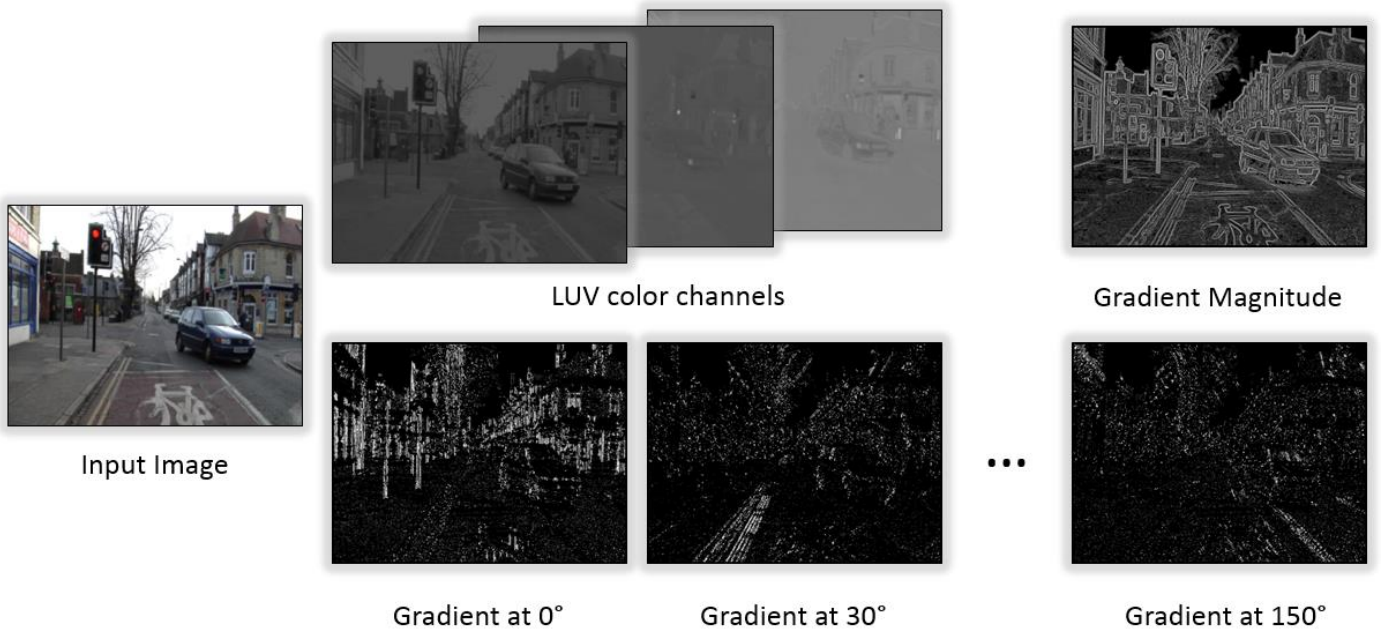
Accelerating Pedestrian Detection

Challenge: Pedestrian detection on mobile devices

- Faster image features
- Faster classification scheme
- State of art accuracy and precision

LUV + HOG Channels

- 10 LUV + HOG image channels [Dollar et al. 2009]:
  - 3 LUV channels
  - 1 gradient magnitude
  - 6 oriented gradient magnitudes



Aggregated channels

ACF approach [Dollar et al. 2014]:

- 4 x 4 pixel aggregation (average computation) => aggregated channels
- Classsification features: simple pixel lookups
- Classifier: boosted two-level decision trees (2048)

$\Rightarrow$ State of art detection at 30 FPS on CPU

Proposed solution:

- Multiresolution features from multiple aggregations:
  - 2 x 2 cells
  - 4 x 4 cells
  - 8 x 8 cells

$\Rightarrow$ 30 aggregated channels

Multiscale detection

Proposed approach:

- 8 pedestrian models: 64, 72, 80, 88, 96, 104, 112, 120 pixel height
- 3 image scales: 1, ½, ¼

⇒ 24 detection scales

Implementation details

Feature computation:

- Lookup tables for: LUV, gradient magnitude and orientation
- Larger aggregation computed from smaller aggregation
- No need for integral images
- No need for approximations for intermediate scales

Classification:

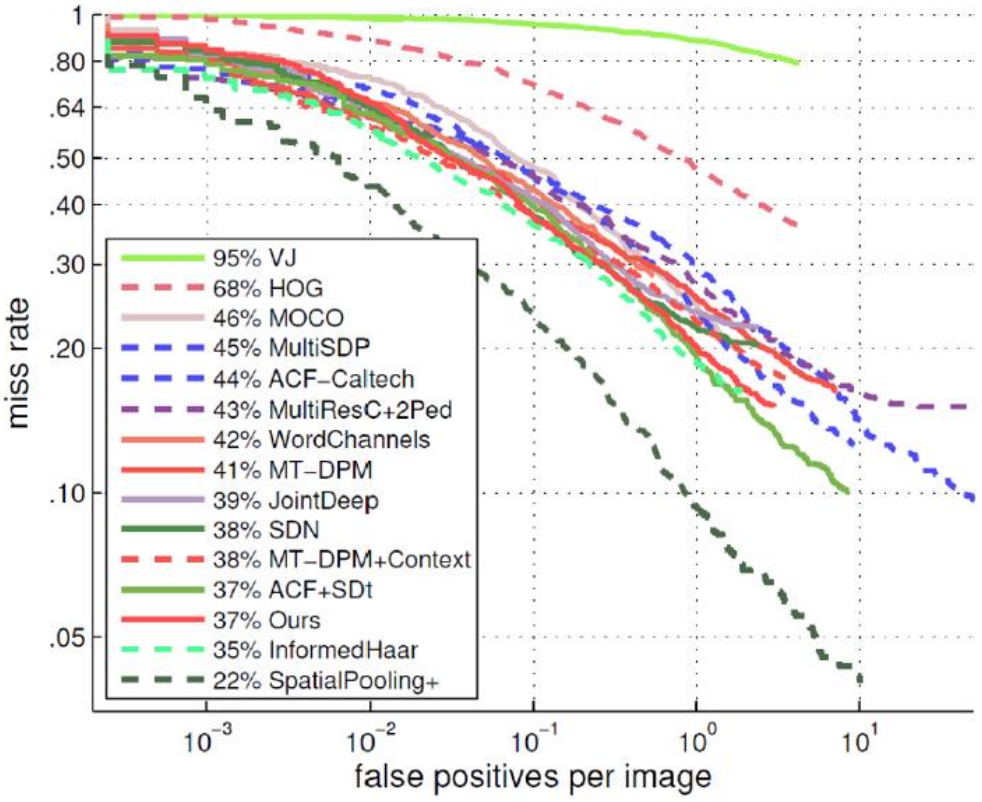
- Prediction using soft-cascade:
  - stop when the classification cost drops below -1
  - 90% rejection after only 32 WLs
- **Early NMS**
  - It is time consuming to evaluate all WLs for overlapping dets.

=> Detection at over 100 FPS on CPU

Validation

Caltech pedestrian detection benchmark – reasonable (2015) :

- **37 %** log-average miss rate for [10^{-2}, 10^{0}] FPPI precision range at **105 FPS**



Approach	Miss rate	FPS
HOG	68.46%	0.05
FPDW	57.40%	2.6
ChnFtrs	56.34%	0.2
CrossTalk	53.88%	14
ACF-Caltech	44.22%	30
WordChannels	42.30%	17
SDN	37.87%	10
SquaresChnFtrs	34.81%	1
InformedHaar	34.60%	0.62
LDCF	24.90%	2
SpatialPooling+	21.89%	0.5
Ours	37.33%	105

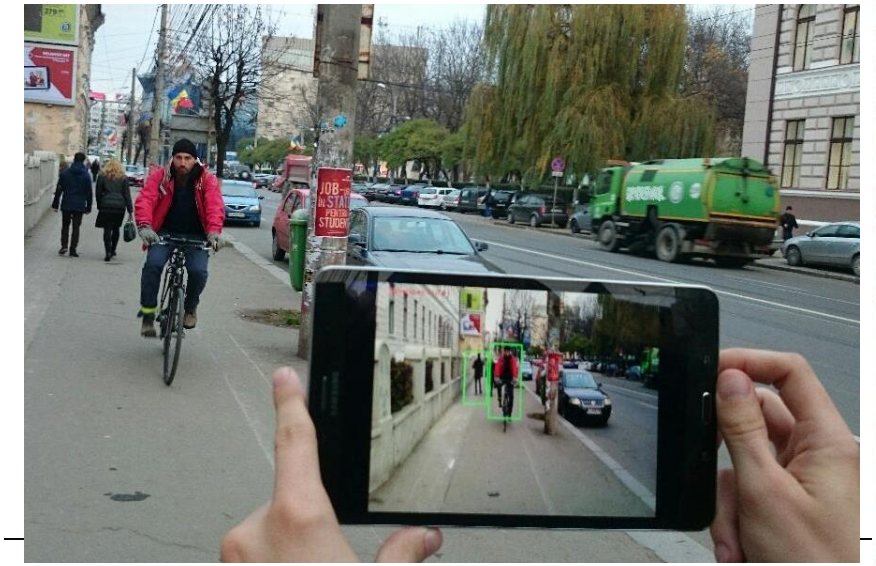
Porting to mobile platforms

The proposed solution was ported and tested on android based mobile devices:

- Samsung Galaxy Tab Pro T325 tablet (Quad-core 2.3 GHz Krait 400 CPU)
- Sony Xperia Z1 smartphone (Quad-core 2.2 GHz Krait 400 CPU)

Detection at: **8 FPS** for pedestrians with heights above 50 pixels

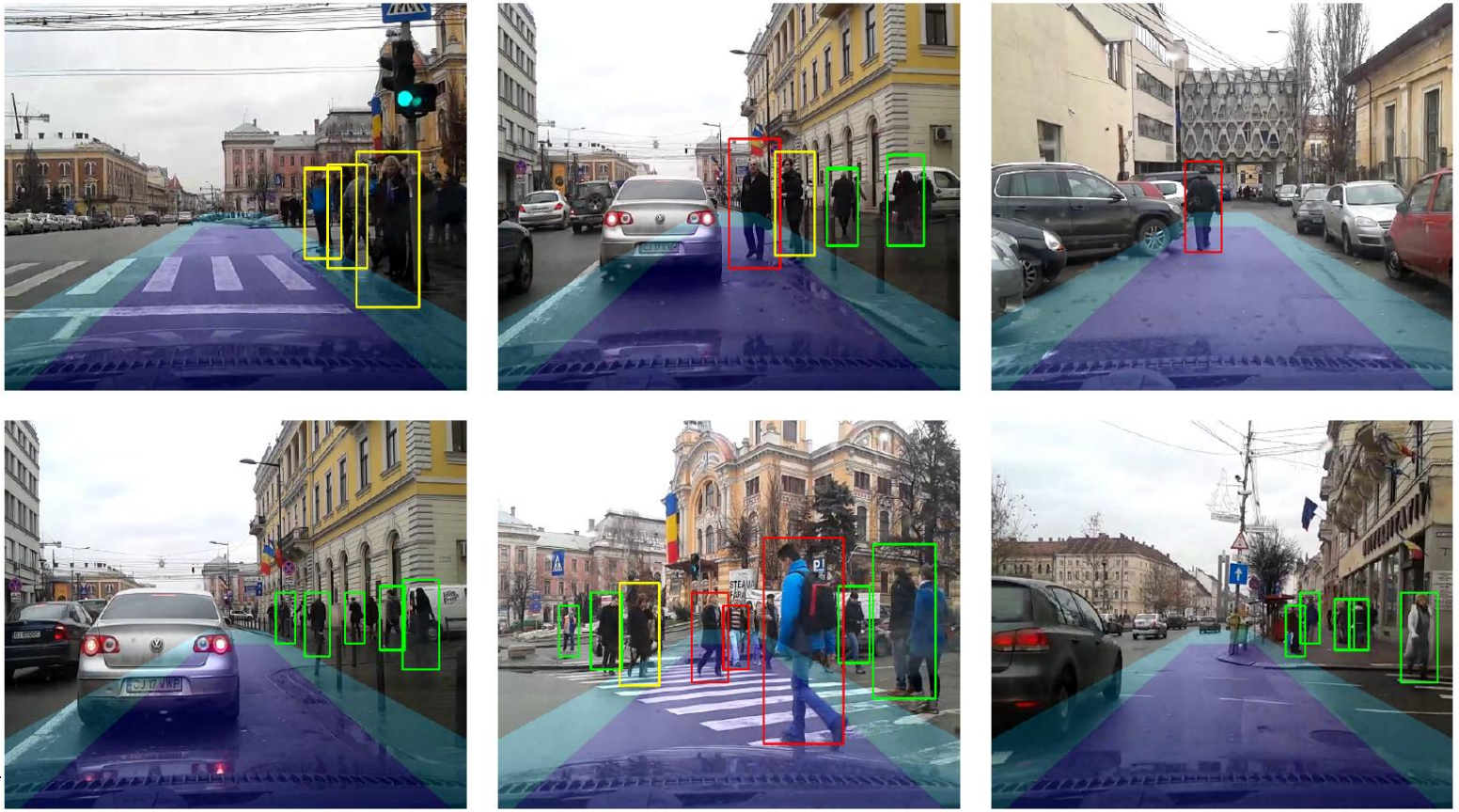
20 FPS for pedestrians with heights above 100 pixels



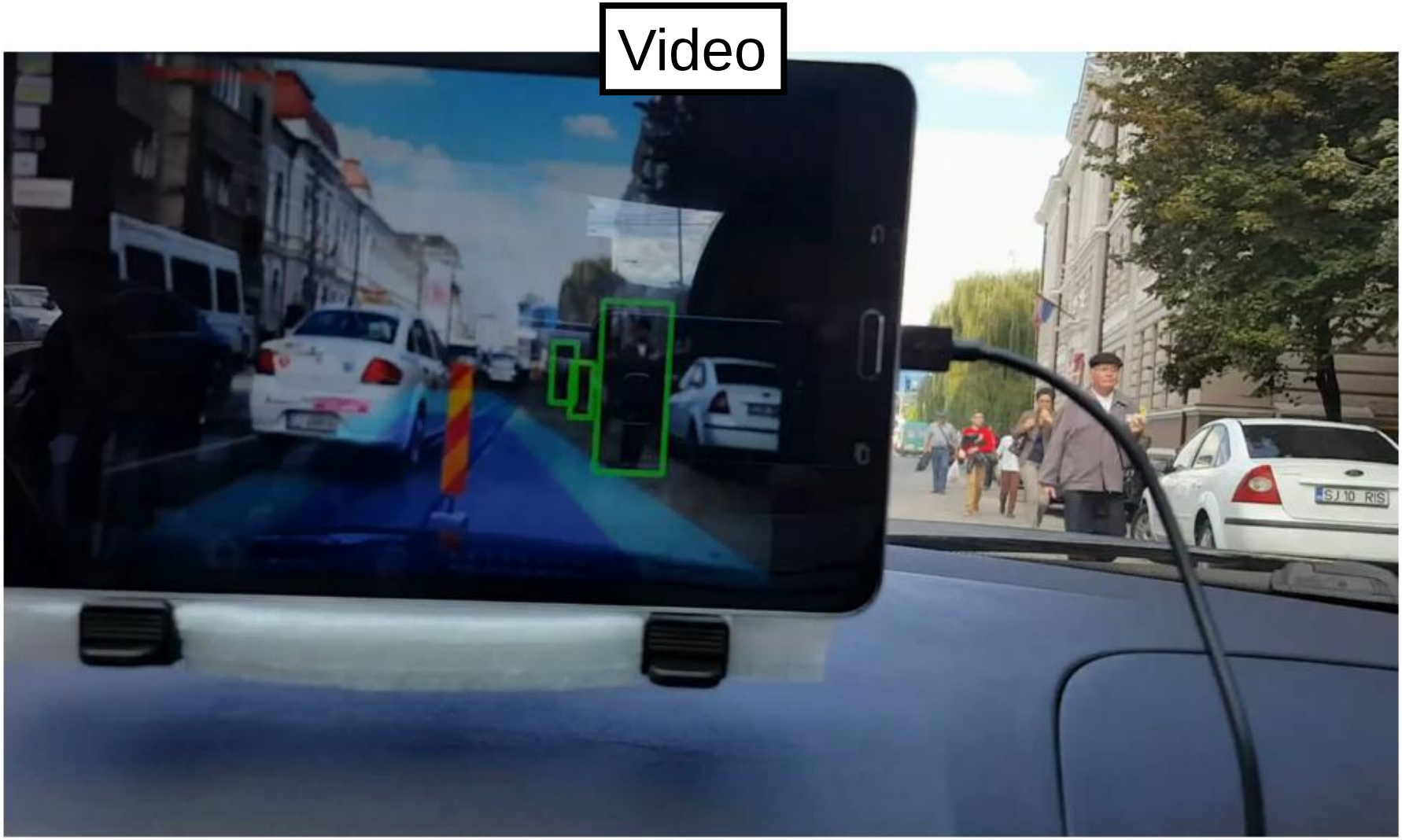
Porting to mobile platforms

Driver assistance application:

- Visual and audio warning when a pedestrian is detected in the front



Demo Application



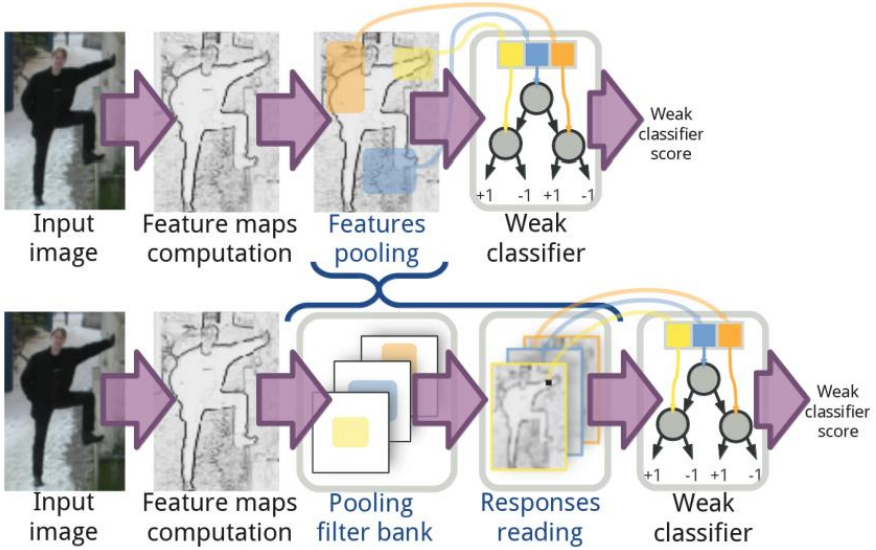
Real-time scene perception

Challenge: real-time perception for autonomous driving

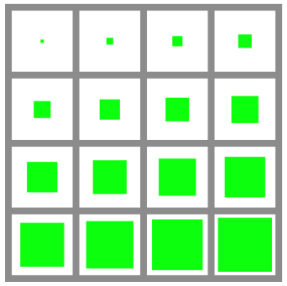
- Need for more powerful features and classification scheme
- Exploitation of multisensorial perception
- Keep computational costs relatively low

Filtered Channels

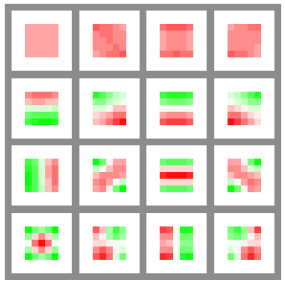
Filtering layer over LUV + HOG channels [Zhang et al. 2015]:



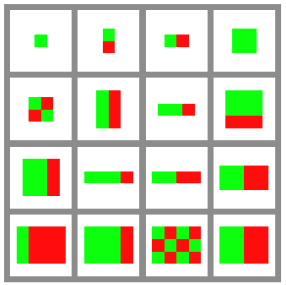
SquaresChntrs Filters



LDCF8 Filters



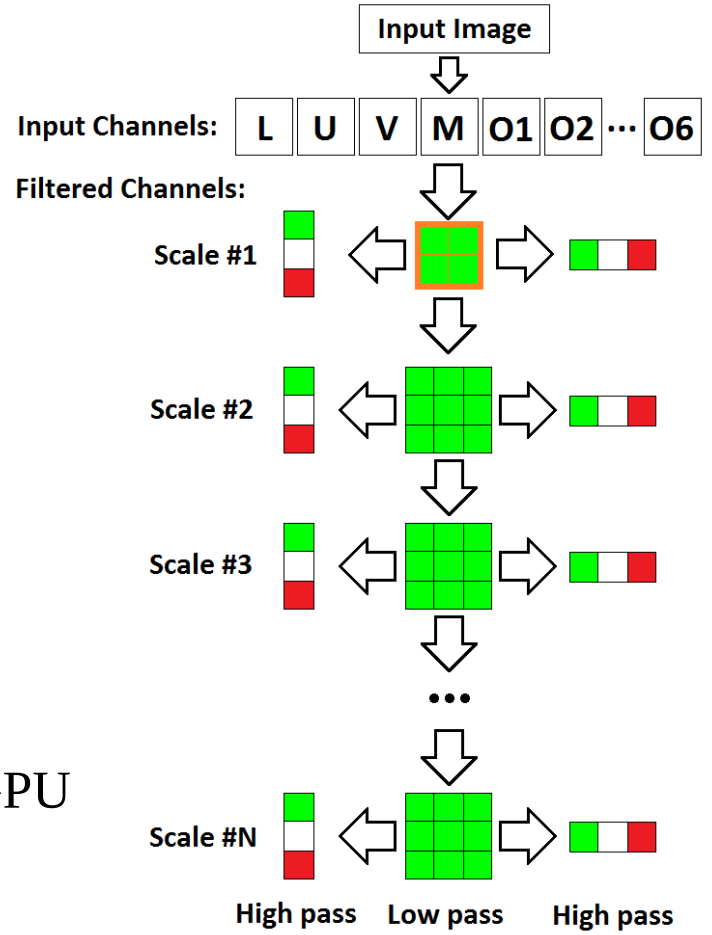
Checkerboards Filters



Multiresolution Filtered Channels

- **Multiresolution** filtering scheme:
  - **Low pass** and **high pass** filters
  - Applied iteratively at **multiple scales**
  - 7 scales => (5 x 3) x 10 = 150 channels

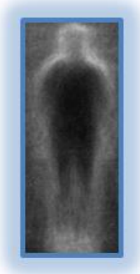
- **Efficient implementation**:
  - **< 3 ms** for a 640 x 480 pixel image on GPU

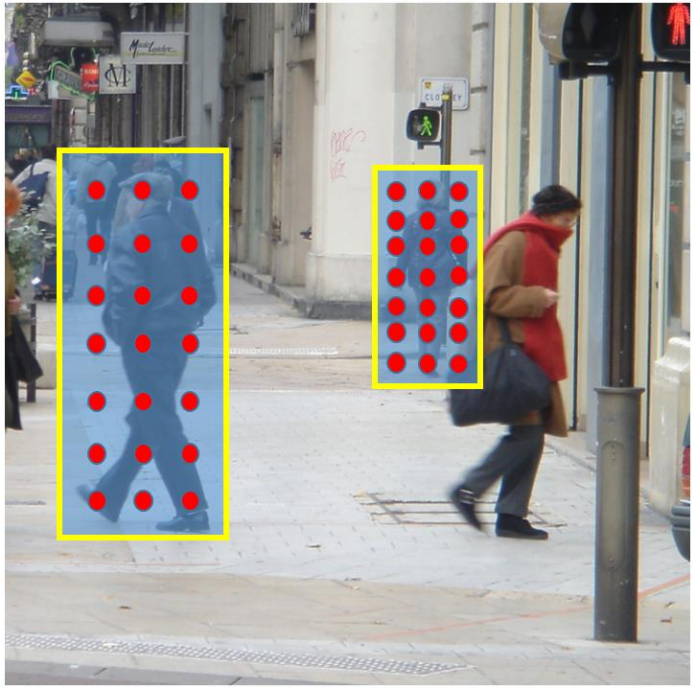


Multiscale Detection

Multiscale sliding window :

- **Single** image feature scale
- **Single** pedestrian classifier model
- Feature sampling adapted to window size



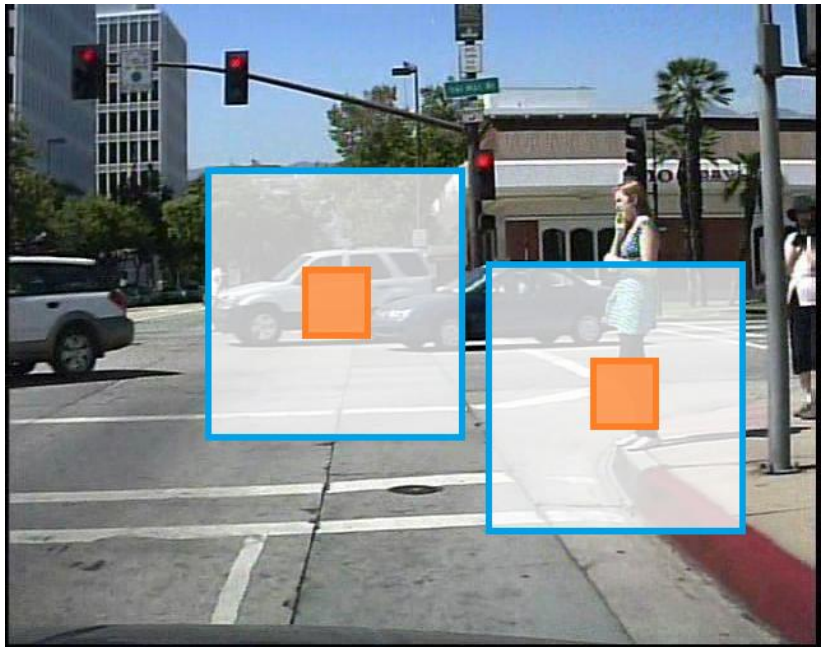


=> Full detection at over 50 FPS

Semantic Segmentation

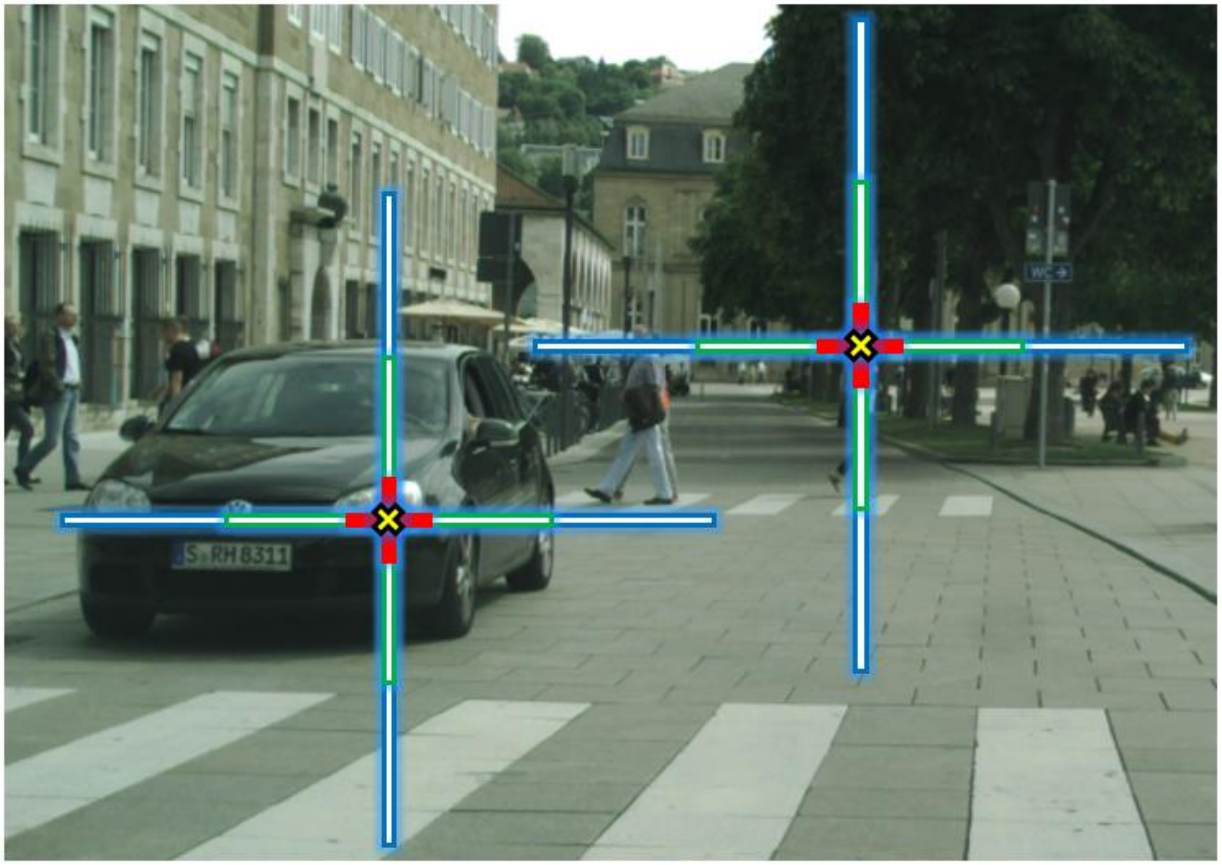
Similar classification scheme for pixels:

- Boosting over Multiresolution Channel features
- **Short** range features => **local structure** - dense sampling
- **Long** range features => **context** - sparse sampling

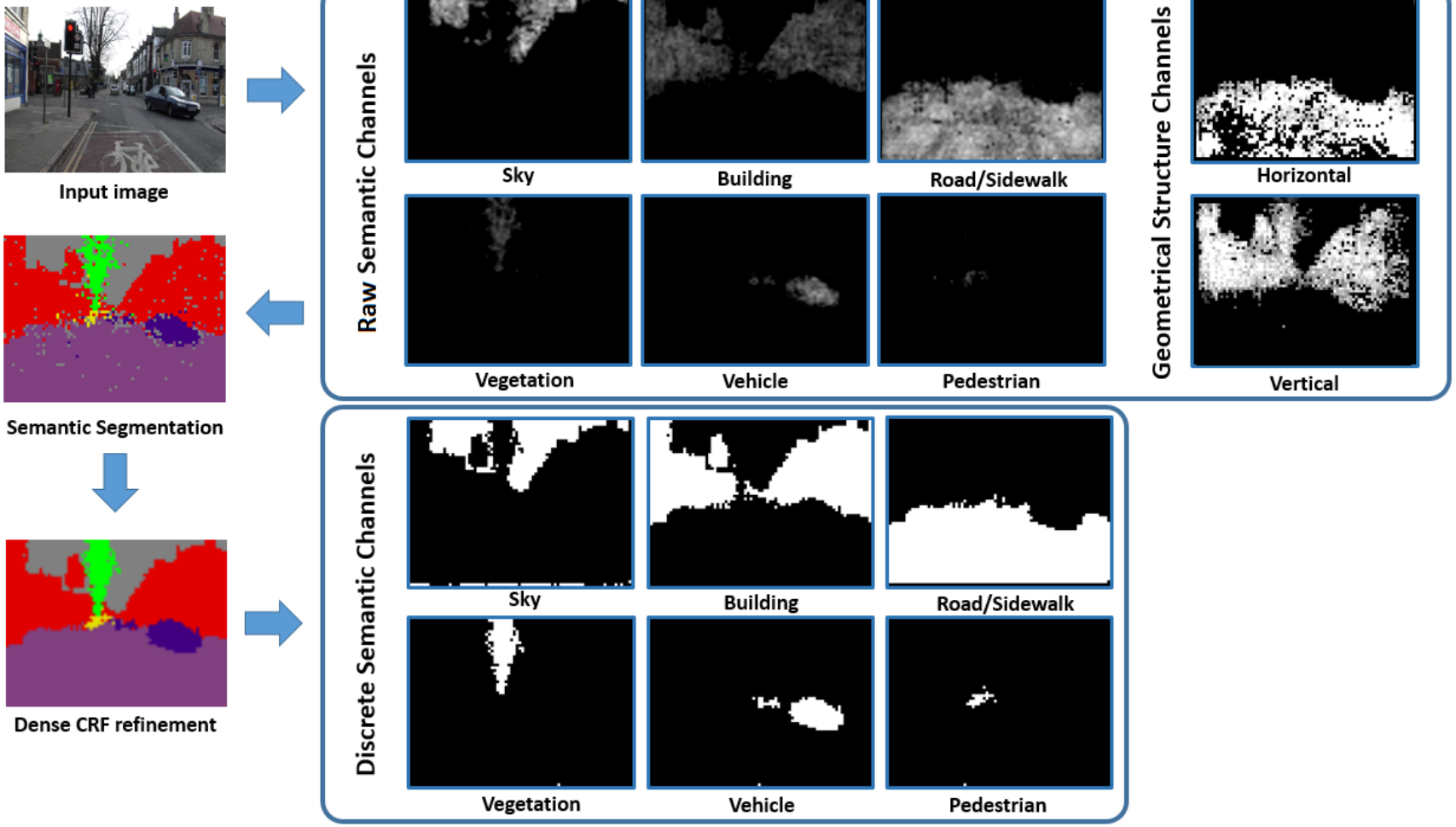


Semantic Segmentation

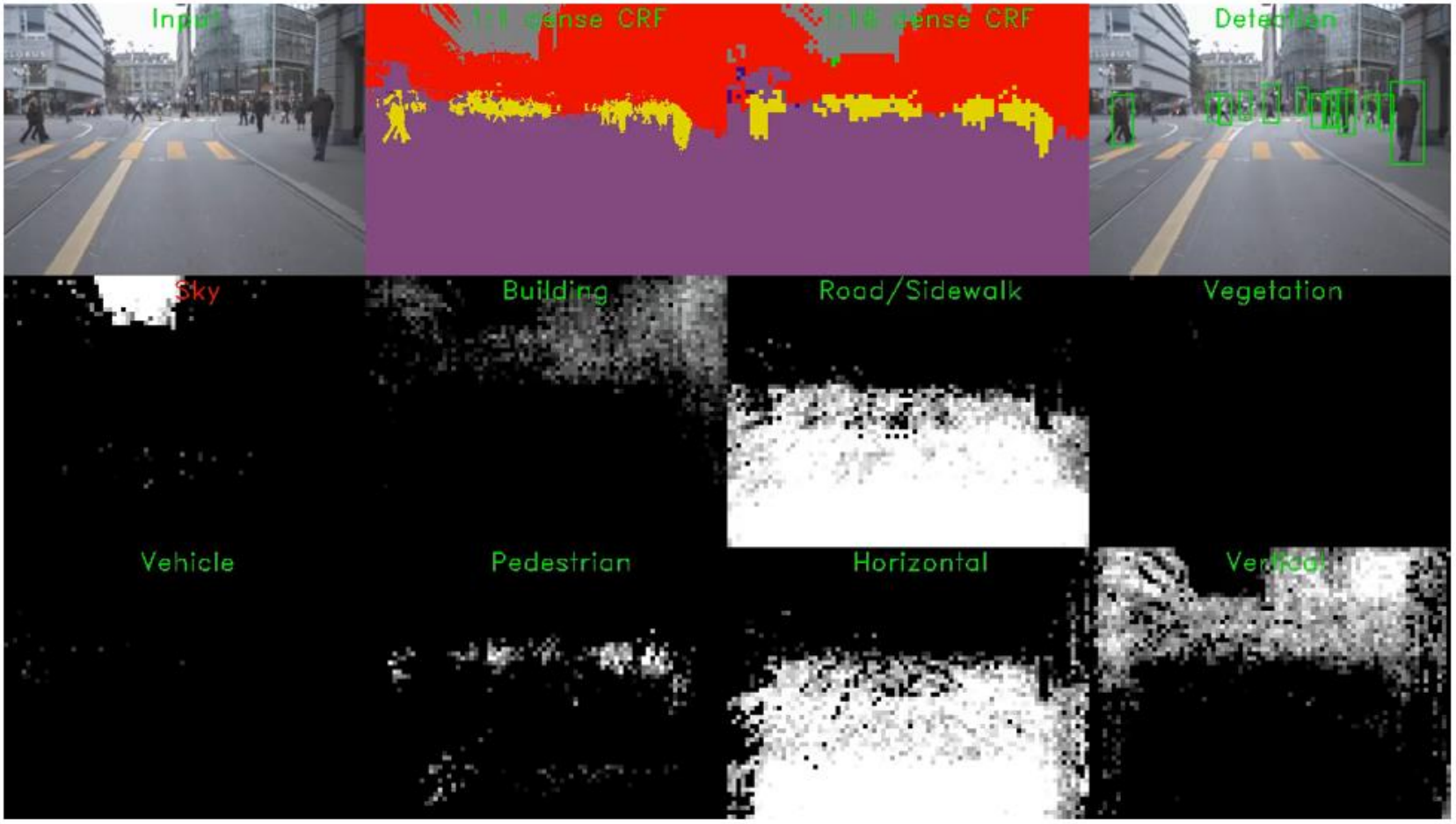
Simplified multi-range classification features (linear sampling):



Semantic Channels for Detection



Detection using MRCF + SemanticCF



Computational costs

Average execution times for different steps (GPU / CPU)

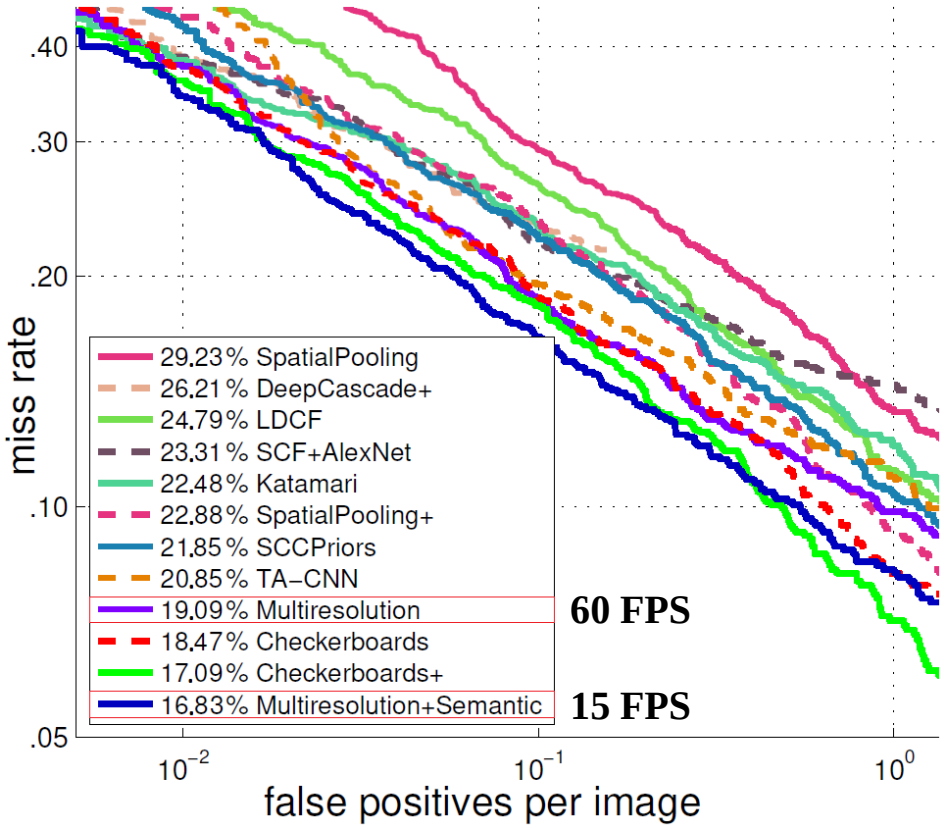
- 210 filtered channel computation: 2 ms / 21 ms
- 8 semmantic channel prediction: 22 ms / 45 ms
- dense CRF inference: - / 28 ms
- sliding window classifications: 14 ms / 29 ms

Average frame rate for pedestrian detection for a 640 x 480 pixel image:

- 60 FPS on GPU / 20 FPS on CPU with 210 filtered channels
- 15 FPS on GPU / 8 FPS on CPU also with semantic channels

Pedestrian detection evaluation

Caltech pedestrian detection benchmark results:



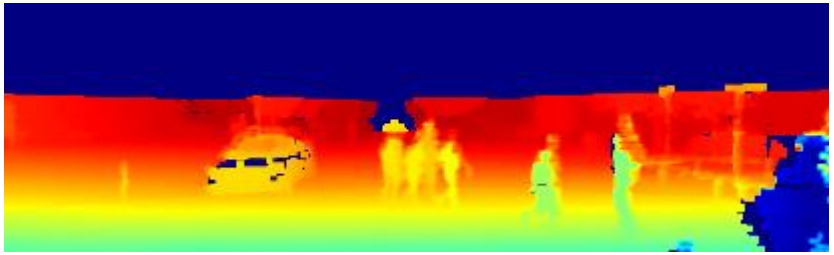
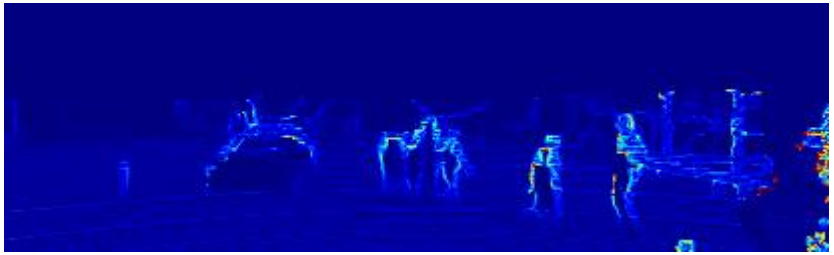
Approach	Miss rate	FPS
FPDW	57.40%	2.6
ChnFtrs	56.34%	0.2
CrossTalk	53.88%	14
Roerei	46.13%	1
ACF-Caltech	44.22%	30
WordChannels	42.30%	16 (GPU)
SDN	37.87%	0.67
FastCF	37.33%	105
LFOV	35.85%	3.6
SquaresChnFtrs	34.81%	1
InformedHaar	34.60%	0.63
SpatialPooling	29.23%	0.13
DeepCascade+	26.21%	15 (GPU)
Ours – Multiresolution	19.09%	20 60 (GPU)
Ours – Multiresolution & Semantic Channels	16.83%	8 15 (GPU)

(2016)

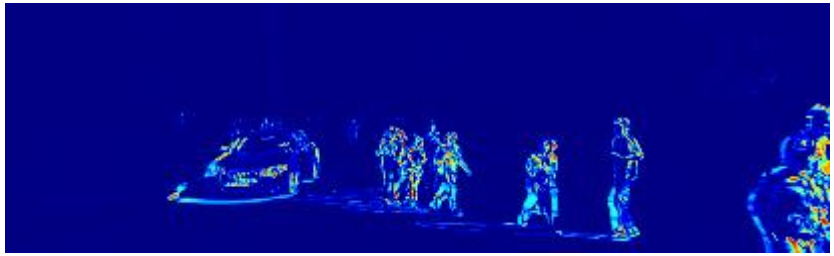
Multimodal Sensorial Input

Color

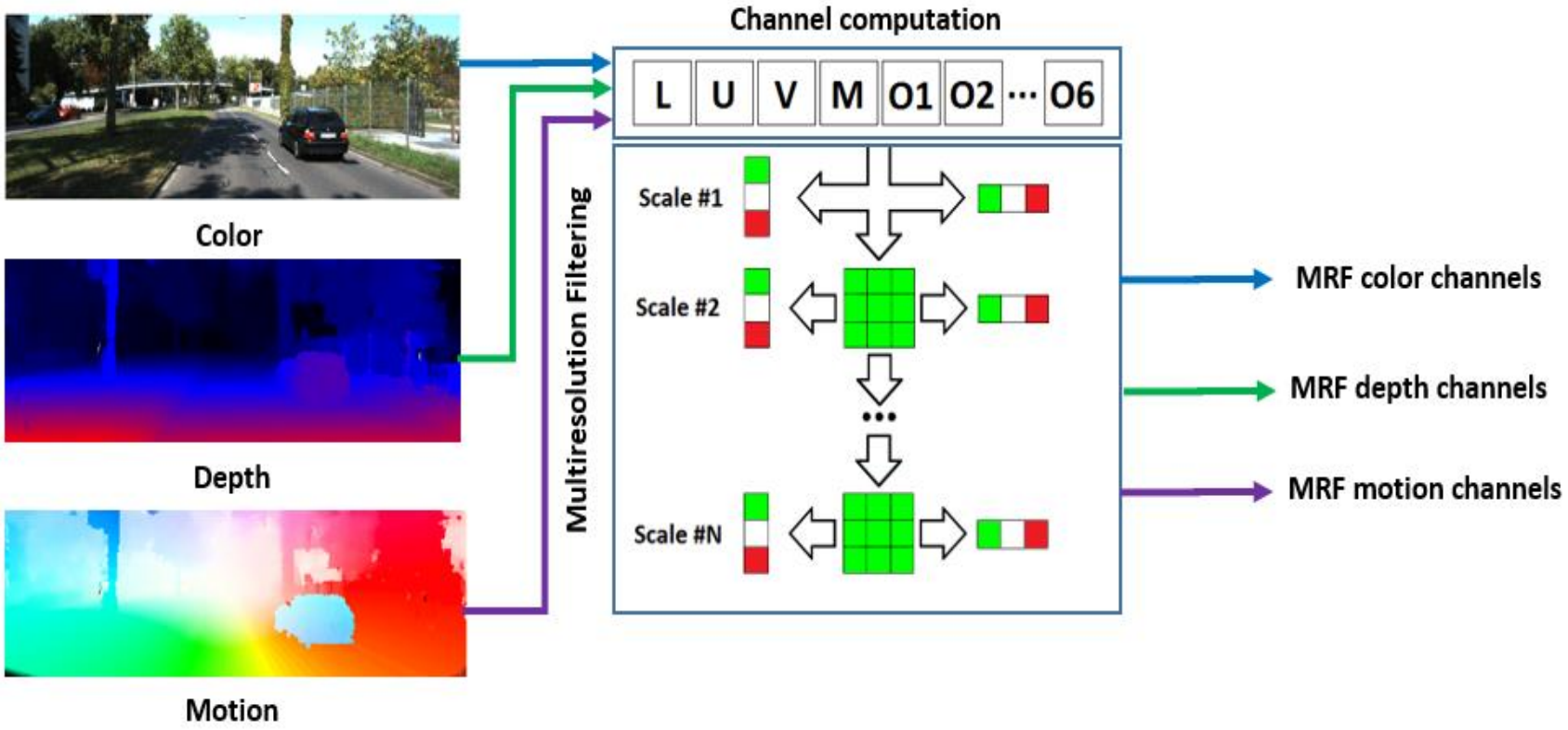

Depth

Motion

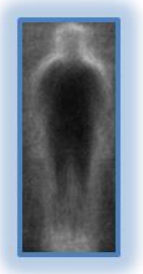


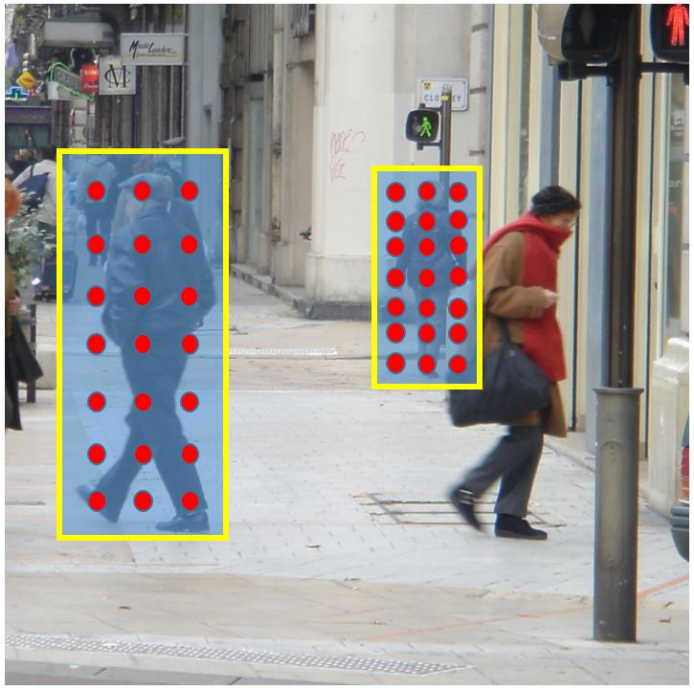

Multimodal Multiresolution Channels



Feature scale correction

One image scale & multiple sliding window scales:





=> **Fast** detection, but the raw channel features are **not scale invariant**

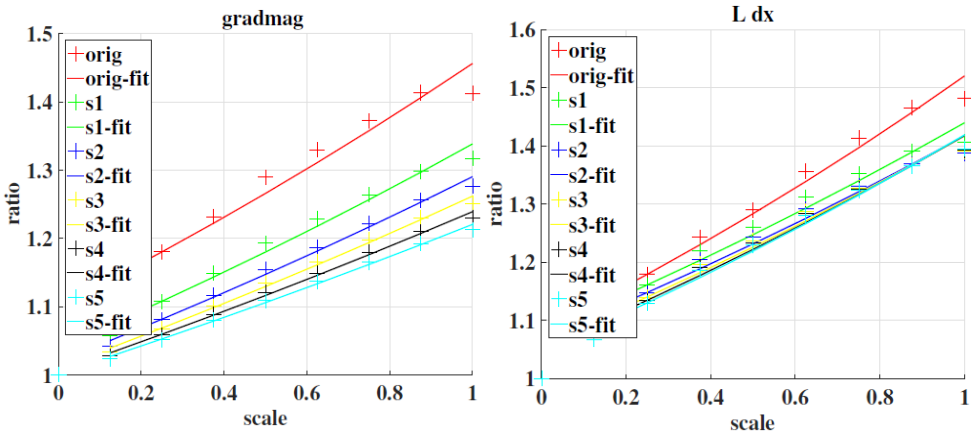
Feature scale correction

Smoothings and derivatives of multimodal intensity can be easily approximated!

Ratio between feature at scale s and at original scale: $r_f(s) = \frac{f(s,x,y)}{f\left(1,\frac{x}{s},\frac{y}{s}\right)}$

Theoretical estimation vs Empirical estimation:

- Intensity or smoothing: $r_I(s) = 1$
- Gradient magnitude: $r_M(s) = \frac{1}{s}$
- Vert. or horiz. derivative: $r_{Idx}(s) = \frac{1}{s}$

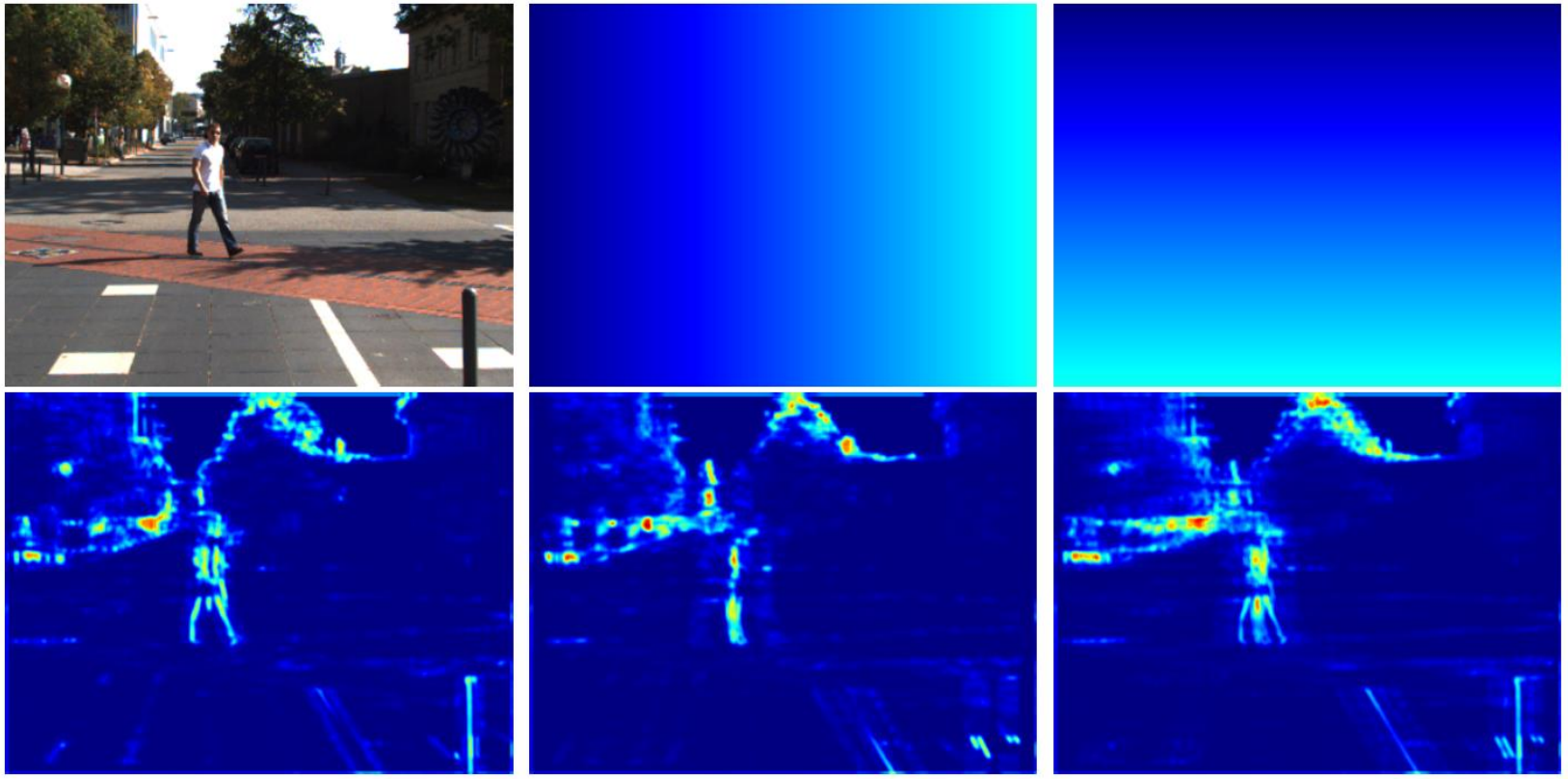


Feature scale correction:

$$f(s,x,y) = r_f(s) \cdot f(1,\frac{x}{s},\frac{y}{s})$$

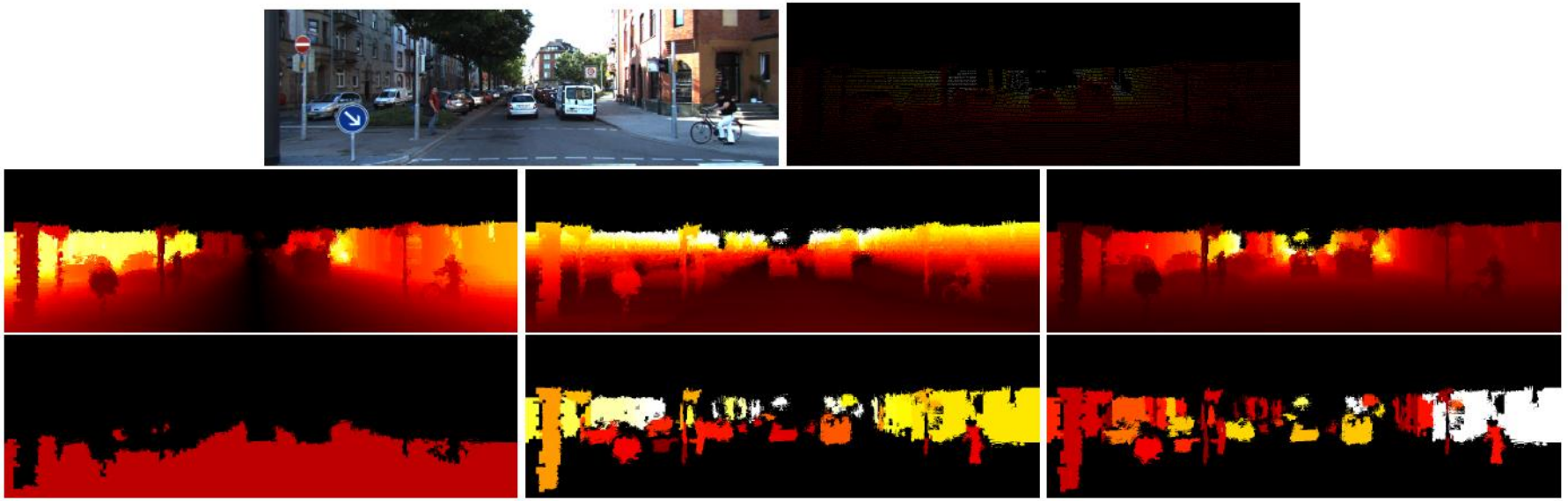
2D context channels

2D spatial and symmetry channels:



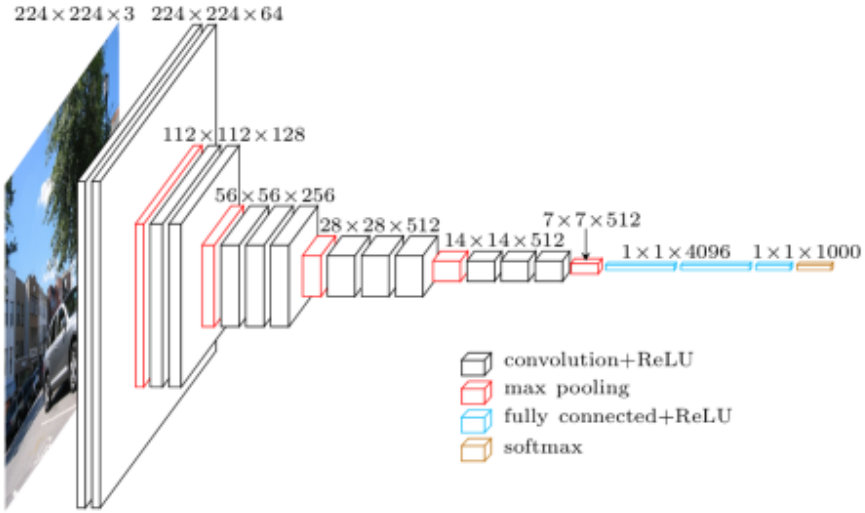
3D Context Channels

- 3D Context channels:
  - Spatial channels: X, Y, Z
  - Ground Plane
  - Geometric channels: height, width, size



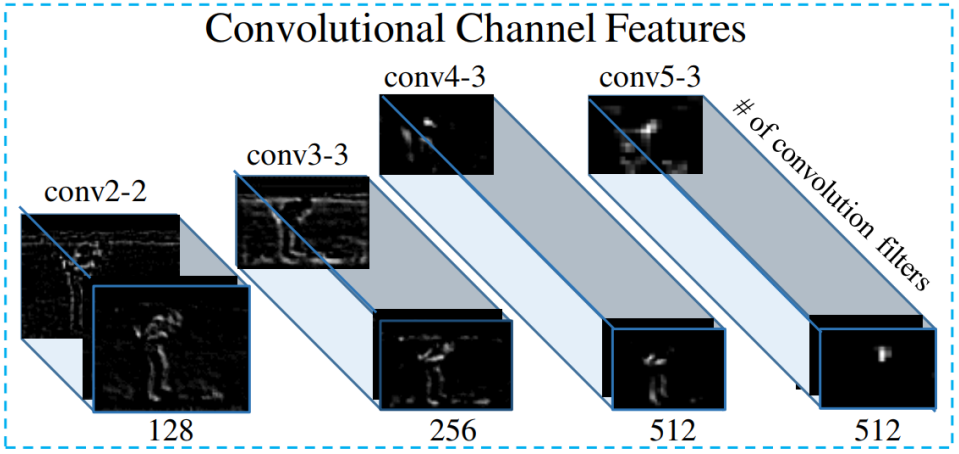
Deep Convolutional Channels

VGG-16 Net [Simonyan and Zisserman 2015]:



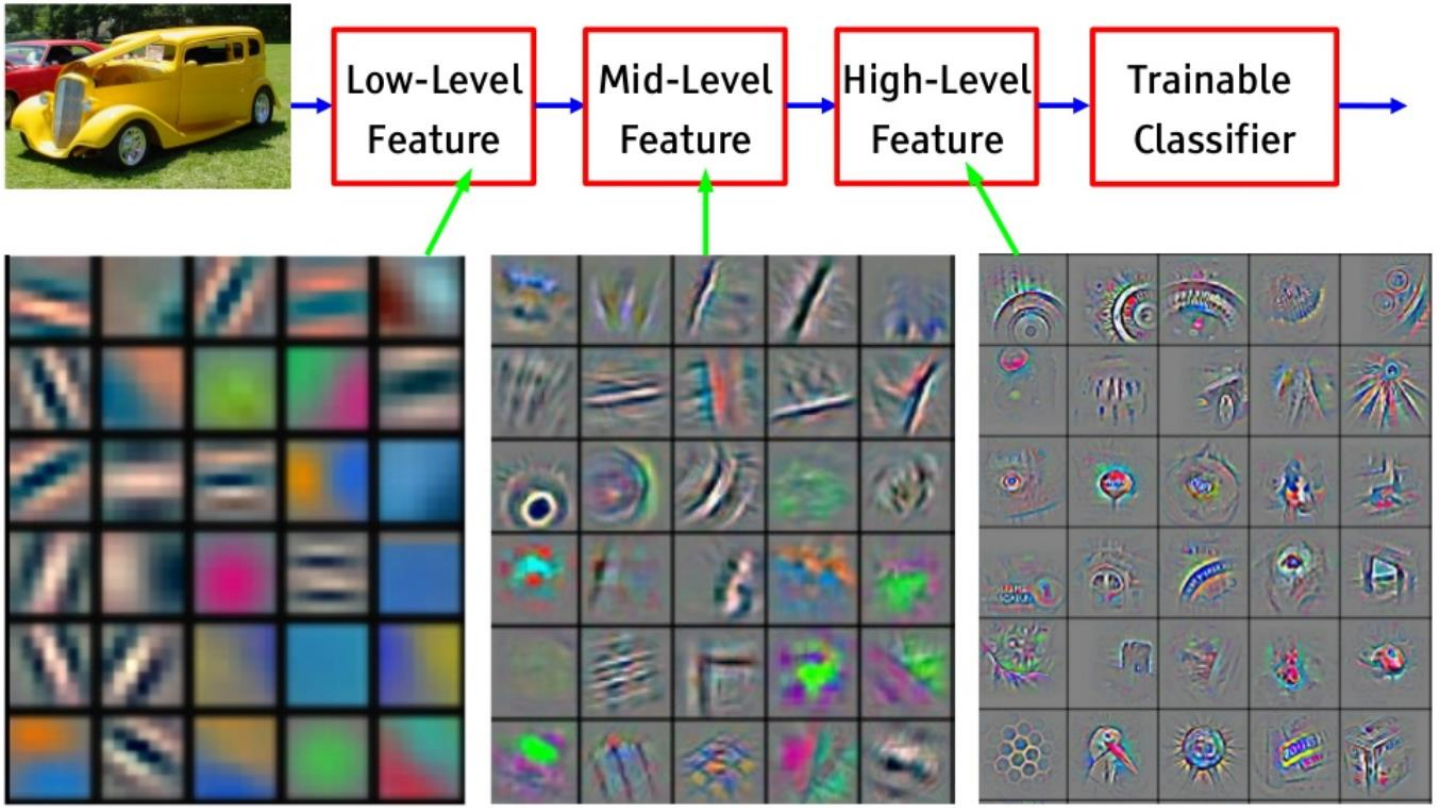
Input Image

[Iqbal et al. 2017]



Deep Convolutional Channels

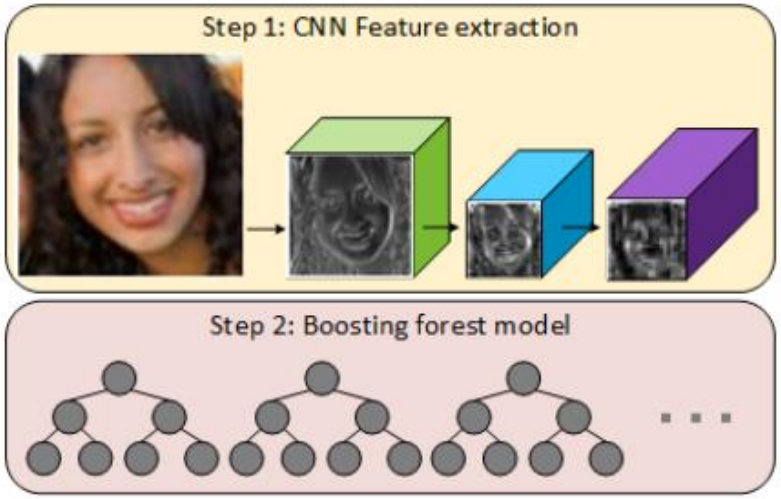
Convolutional net feature visualization [Zeiler & Fergus 2013]



Deep Convolutional Channels

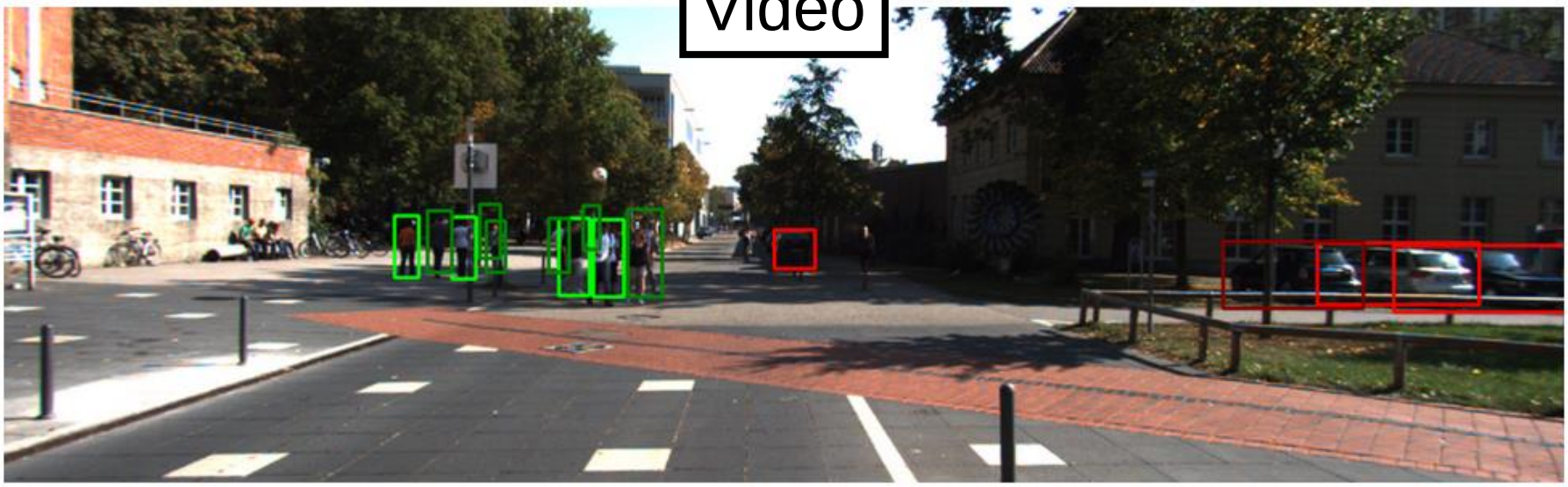
Convolutional channel features [Yang et al. 2015]:

- best results for pedestrian detection using the standard VGG16 pre-trained model
- VGG16 was trained for 2 weeks on ImageNet (over 1 million images, 1000 classes)



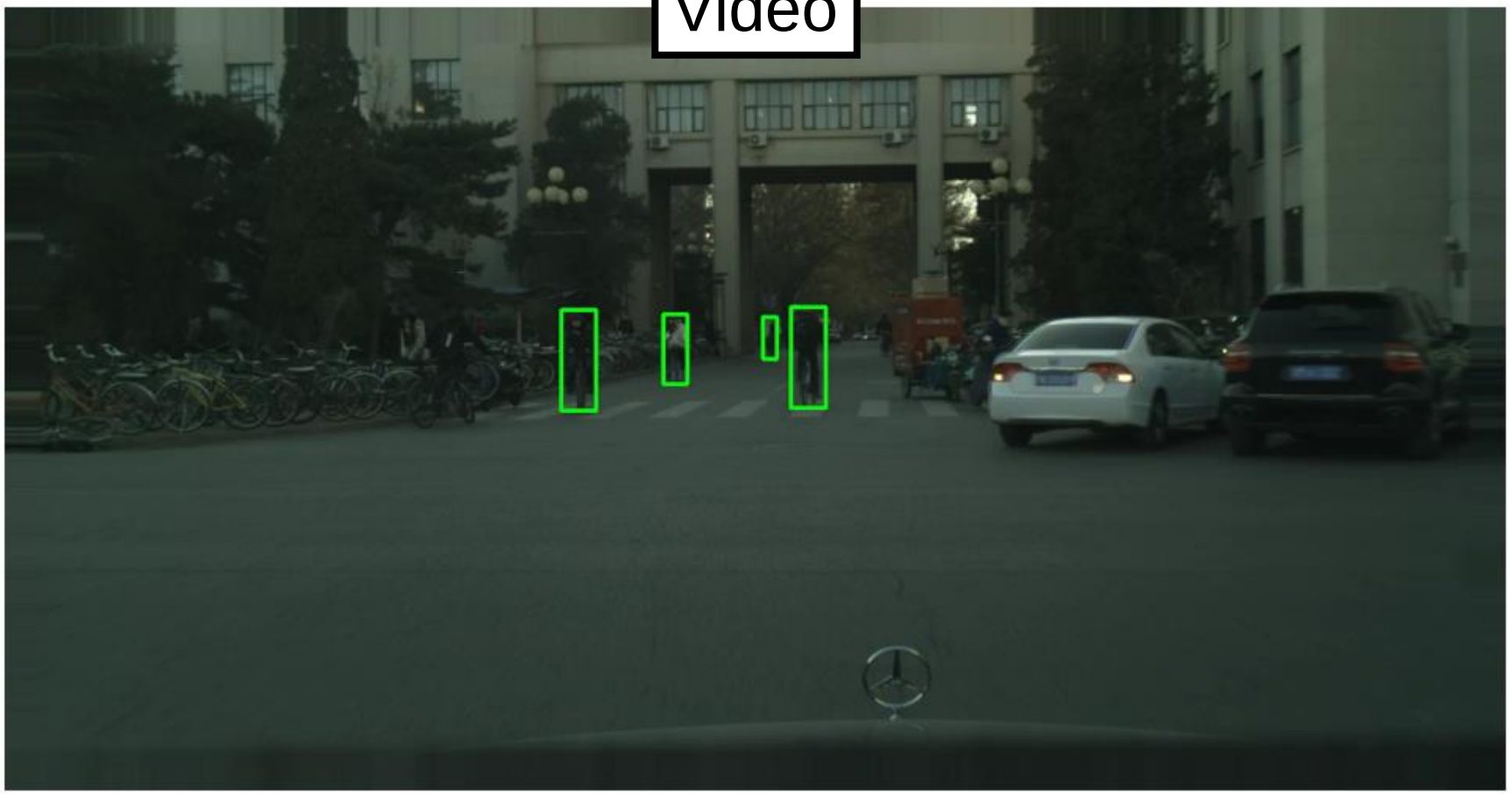
	Output layer	#Output maps	Filter size	#Ds	Miss Rate(%)
ACF	-	10	3	4	**41.22**
LDCF	-	40	7	4	**38.66**
ANet-s1	conv1	96	11	4	61.65
	conv2	256	5	4	51.52
	conv3	384	3	4	**43.73**
	conv4	384	3	4	48.37
	conv5	256	3	4	53.37
VGG-16	conv2-2	128	3	4	53.86
	conv3-3	256	3	4	31.28
	conv4-3	512	3	8	**27.66**
	conv5-3	512	3	16	51.52
VGG-19	conv2-2	128	3	4	51.25
	conv3-4	256	3	4	33.56
	conv4-4	512	3	8	**30.17**
	conv5-4	512	3	16	55.55
GNet	conv2	192	3	4	45.06
	icp1	256	-	8	38.44
	icp2	480	-	8	**31.66**
	icp3	512	-	16	35.99
GNet-s1	conv2	192	3	4	49.39
	icp1	256	-	4	41.85
	icp2	480	-	4	**32.18**
	icp3	512	-	8	32.87

Detection Demo (KITTI)



Pedestrian and vehicle detection using color, motion and depth (LIDAR)

Detection Demo (Tsinghua - Daimler)

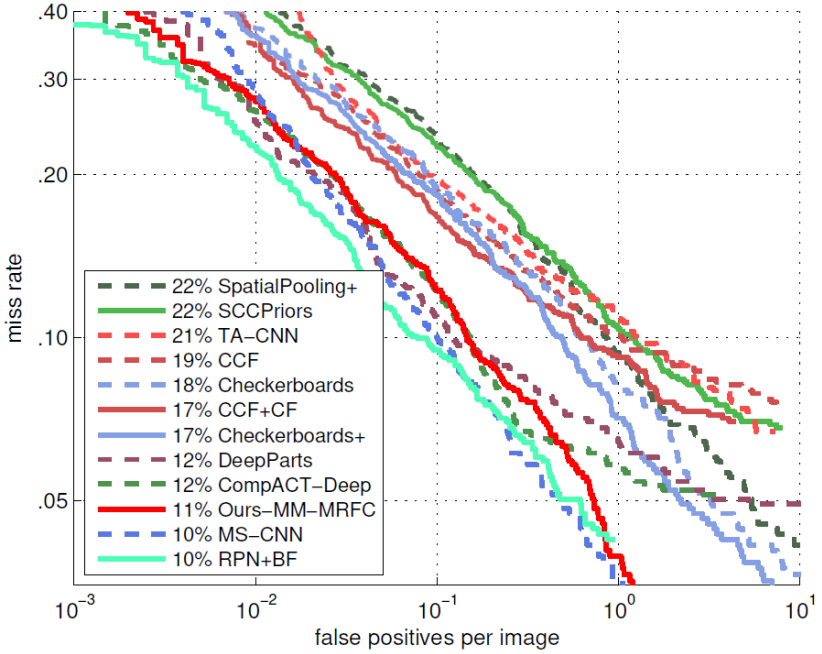


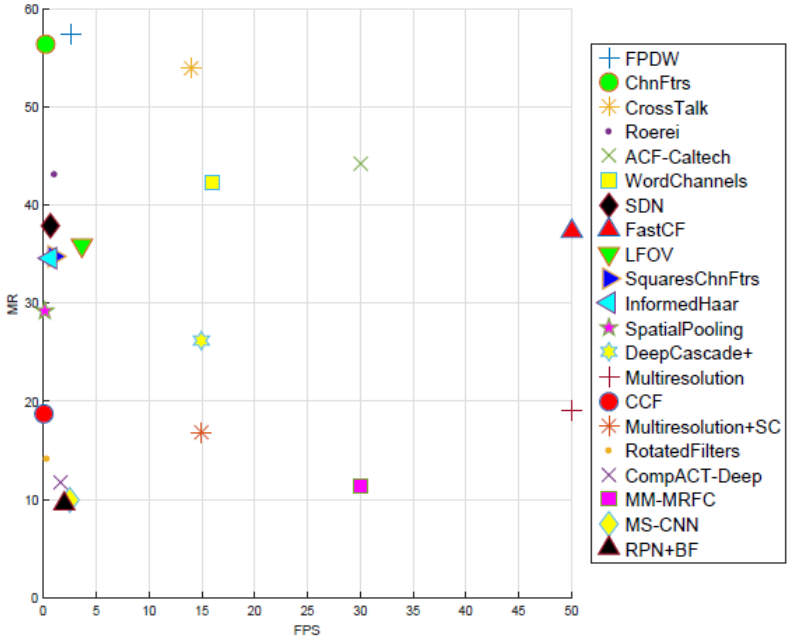
Cyclist detection using color and depth (stereo)

Detection evaluation

Caltech Pedestrian detection benchmark - reasonable:

- 11.41 % avg. MR at 30 FPS
- 9.58 % avg. MR at 25 FPS using deep conv. chnl. features





(2017)

Detection evaluation

Feature evaluation for pedestrian detection:

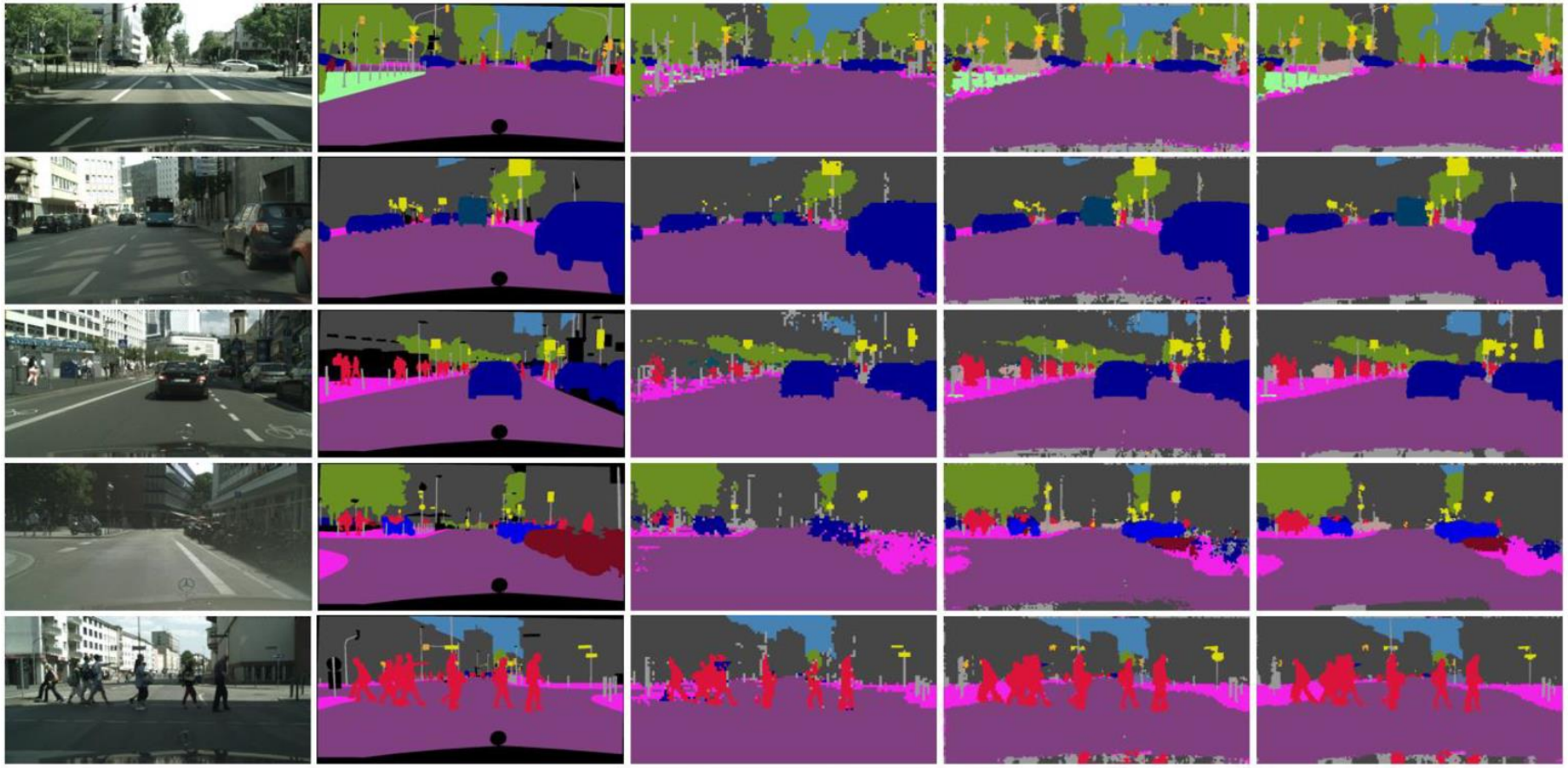
Caltech

Channel Type		Caltech MR - reasonable -
Color	MRFC no SC	24.46
	MRFC E-SC	22.69
	MRFC T-SC	22.84
	+ 2D spatial	20.80
	+ 2D symmetry	18.26
Motion	+ SDt	17.29
	+ MM-MRFC	16.11

KITTI (val)

Context Type		KITTI AP		
		Easy	Moderate	Hard
Color	MRFC no SC	62.84	59.98	51.10
	MRFC	67.14	61.45	52.76
	+ 2D spatial	69.58	63.83	54.83
	+ 2D symmetry	70.28	64.75	55.66
3D stereo	+ 3D spatial	77.88	70.30	60.63
	+ 3D geometric	77.97	70.61	61.47
	+ MRFC	82.53	74.82	65.95
3D LIDAR	+ 3D spatial	77.88	70.93	61.91
	+ 3D geometric	79.92	72.48	63.13
	+ MRFC	84.26	76.34	67.18
Motion	+MRFC	85.25	77.72	68.28

Segmentation results (Cityscapes)



Segmentation results (Cityscapes)

Cityscapes test set - comparison:

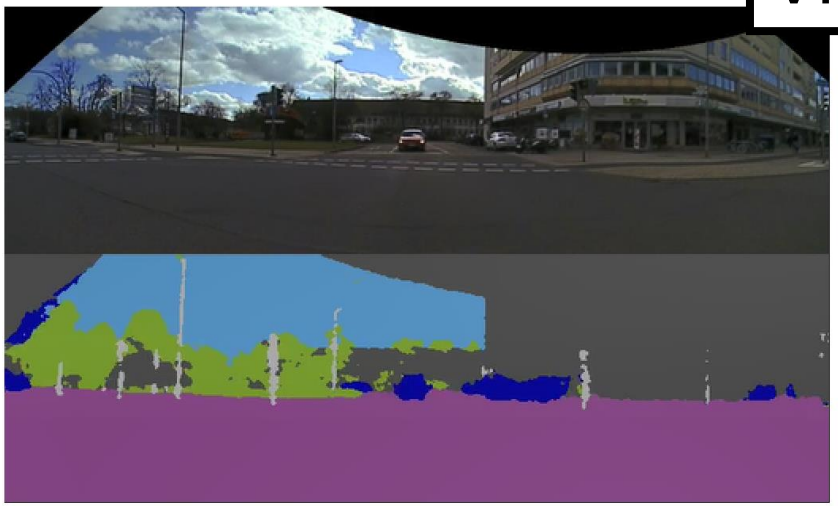
	Class Mean IoU	Category Mean IoU	Runtime [s]
Dilation10	67.1	86.5	4.0
Adelaide	66.4	82.8	35.0
FCN8s	65.3	85.7	0.5
DeepLab LargeFOV StrongWeak	64.8	81.3	4.0
DeepLab LargeFOV Strong	63.1	81.2	4.0
CRFasRNN	62.5	82.7	0.7
SQ	59.8	84.3	0.06
ENet	58.3	80.4	0.013
Segnet basic	57.0	79.1	0.06
Segnet extended	56.1	79.8	0.06
MultiBoost	59.2	81.8	0.25

360 degree semantic perception

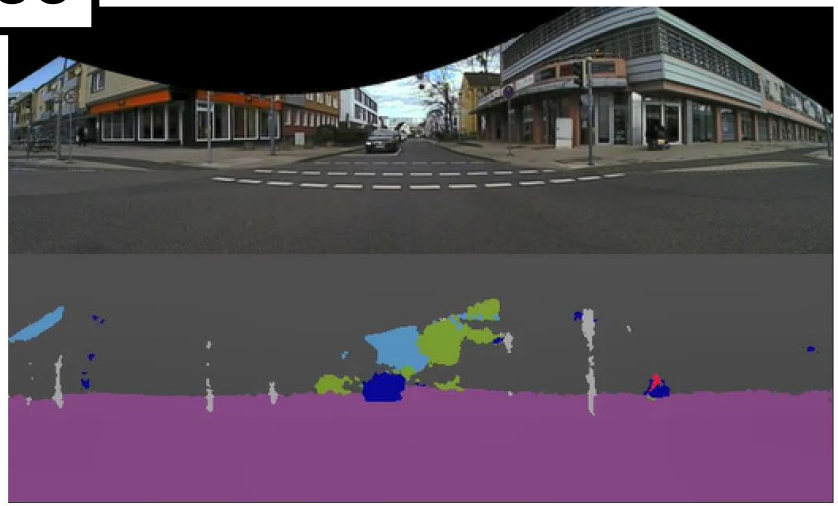
Front

Video

Back



Left



Right

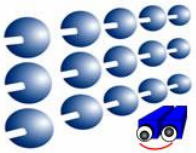
Conclusions

- Channel types:
  - Word channels
  - LUV + HOG:
    - Aggregated channels (single or multiple times)
    - Multiresolution filtered channels (MRFC)
  - Multimodal MRFC
  - 2D & 3D context channels
  - Semantic channels
  - Deep convolutional channels
- Boosting over channel features can be a powerful tool:
  - enables easy fusion of different feature types
  - computational cost friendly
  - easy tuning

Conclusions

More details can be found in:

- A. D. Costea, R. Varga, S. Nedevschi, "Fast Boosting based Detection using Scale Invariant Multimodal Multiresolution Filtered Features", IEEE Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 2017
- A. D. Costea, S. Nedevschi, "Traffic Scene Segmentation based on Boosting over Multimodal Low, Intermediate and High Order Multi-range Channel Features", IEEE Intelligent Vehicles Symposium (IV), Redondo Beach, USA, 2017
- A. D. Costea, S. Nedevschi, "Semantic Channels for Fast Pedestrian Detection", IEEE Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016
- A. D. Costea, S. Nedevschi, "Fast Traffic Scene Segmentation using Multi-range Features from Multi-resolution Filtered and Spatial Context Channels", IEEE Intelligent Vehicles Symposium (IV), Gothenburg, Sweden, 2016
- A. D. Costea, A. V. Vesa, S. Nedevschi, "Fast Pedestrian Detection for Mobile Devices", IEEE Intelligent Transportation Systems Conference (ITSC), Las Palmas de Gran Canaria, Spain, 2015
- A. D. Costea, S. Nedevschi, "Word channel based multiscale pedestrian detection without image resizing and using only one classifier", IEEE Computer Vision and Pattern Recognition, (CVPR), Columbus, USA, 2014
- A. D. Costea, S. Nedevschi, "Multi-class segmentation for traffic scenarios at over 50 fps", IEEE Intelligent Vehicles Symposium (IV), Dearborn, USA, 2014



Thank you for your attention!

Questions?