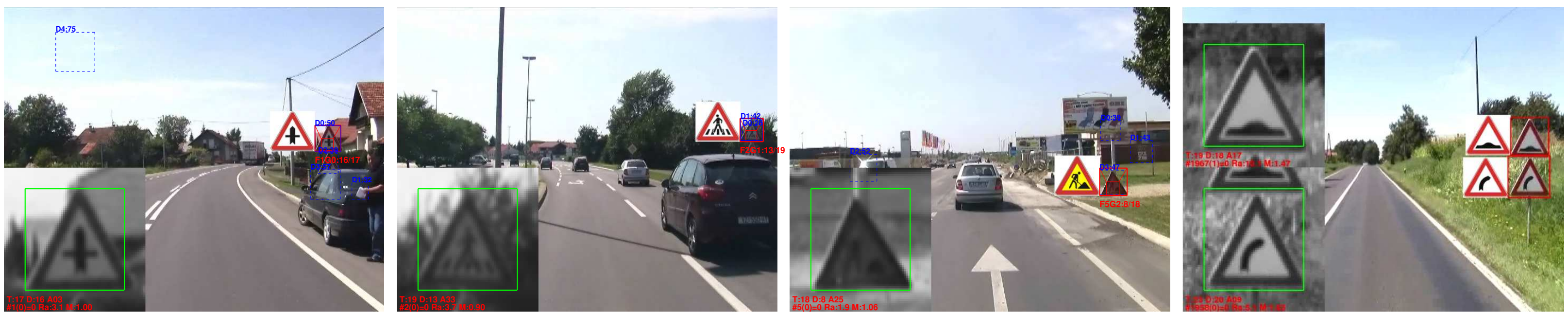
Recent advances in traffic sign detection



Siniša Šegvić

University of Zagreb

Graz, 1st December 2011.

Agenda

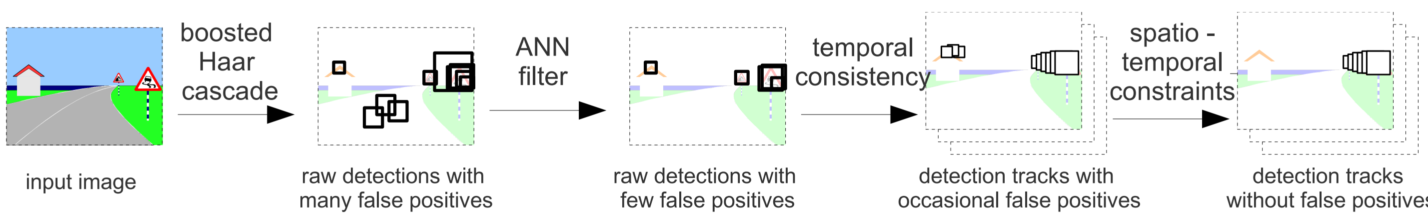
- Introduction: motivation and challenges

Agenda

- Introduction: motivation and challenges
- Assumptions and datasets

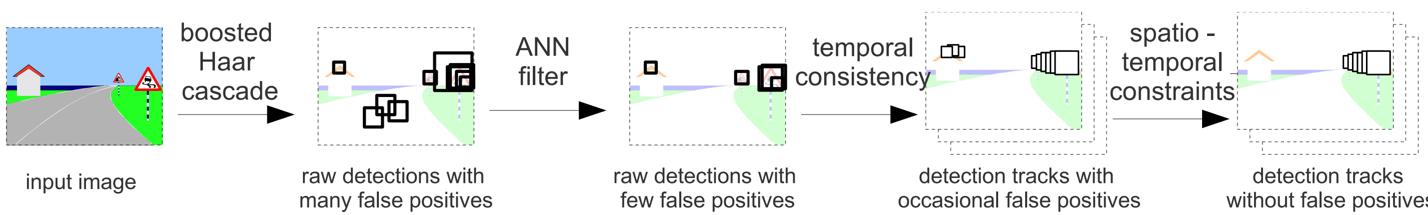
Agenda

- Introduction: motivation and challenges
- Assumptions and datasets
- The devised detection pipeline:
  - baseline detection by boosted Haar cascades
  - additional heterogeneous cascade stage for improved *precision*
  - enforcing temporal consistency to improve *localization accuracy*
  - enforcing learned contextual constraints



Agenda

- Introduction: motivation and challenges
- Assumptions and datasets
- The devised detection pipeline:
  - baseline detection by boosted Haar cascades
  - additional heterogeneous cascade stage for improved *precision*
  - enforcing temporal consistency to improve *localization accuracy*
  - enforcing learned contextual constraints



- Future challenges:
  - simultaneous detection of different sign classes
  - generic detection of table-like objects

Introduction: motivation

Why would we like to detect traffic signs in images?

- on-board applications: driver assistance, autonomous driving
- off-board applications: road safety inspection

Introduction: motivation

Why would we like to detect traffic signs in images?

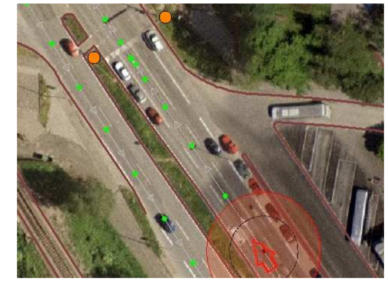
- on-board applications: driver assistance, autonomous driving
- off-board applications: road safety inspection

Why do we need road safety inspection?

- crucial for detecting safety issues of a road in operation
- in practice, the inspection mainly concerns anomalies of the traffic control devices:
  - damaged, covered, worn-out or stolen signs
  - erased or incorrectly painted surface markings
- safety determined by **assessment frequency**

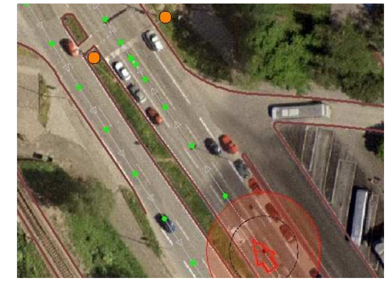
INTRODUCTION: MOTIVATION (2)

In current commercial practice inspection is performed by expensive and subjective human experts



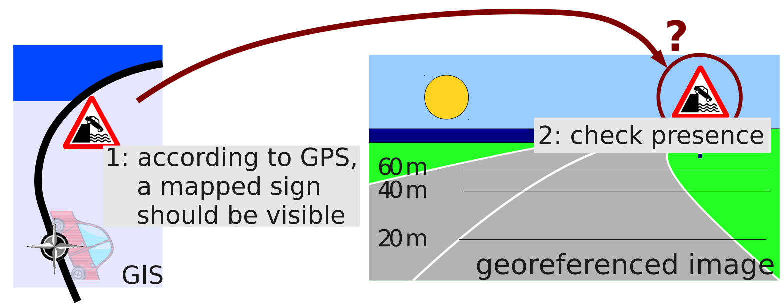
Introduction: motivation (2)

In current commercial practice inspection is performed by expensive and subjective human experts



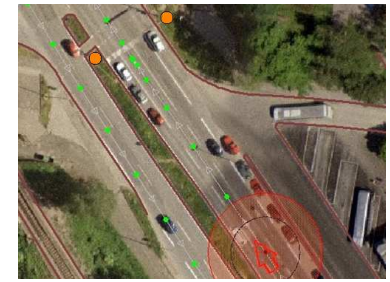
An innovation opportunity: automate inspection of the elements of traffic infrastructure in order to achieve better service for less money

assessment: are the mapped elements present in a recent geo-referenced video?



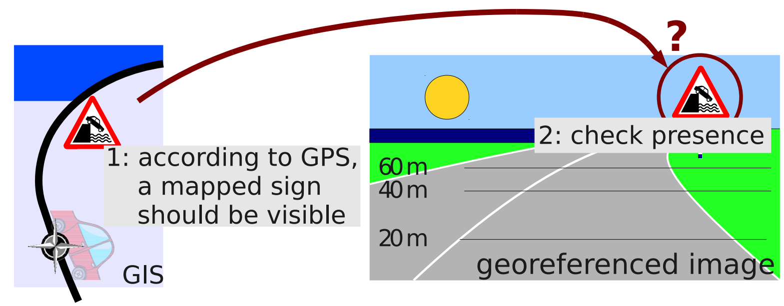
Introduction: motivation (2)

In current commercial practice inspection is performed by expensive and subjective human experts

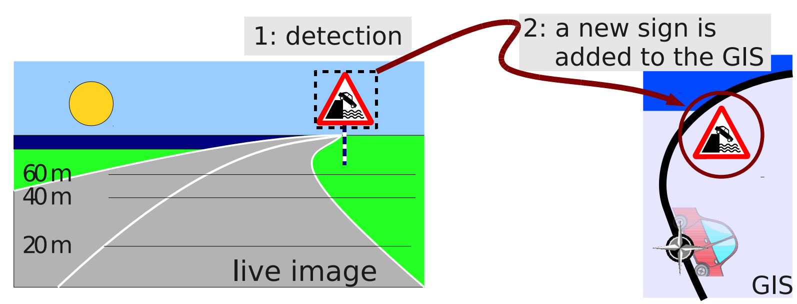


An innovation opportunity: automate inspection of the elements of traffic infrastructure in order to achieve better service for less money

assessment: are the mapped elements present in a recent geo-referenced video?

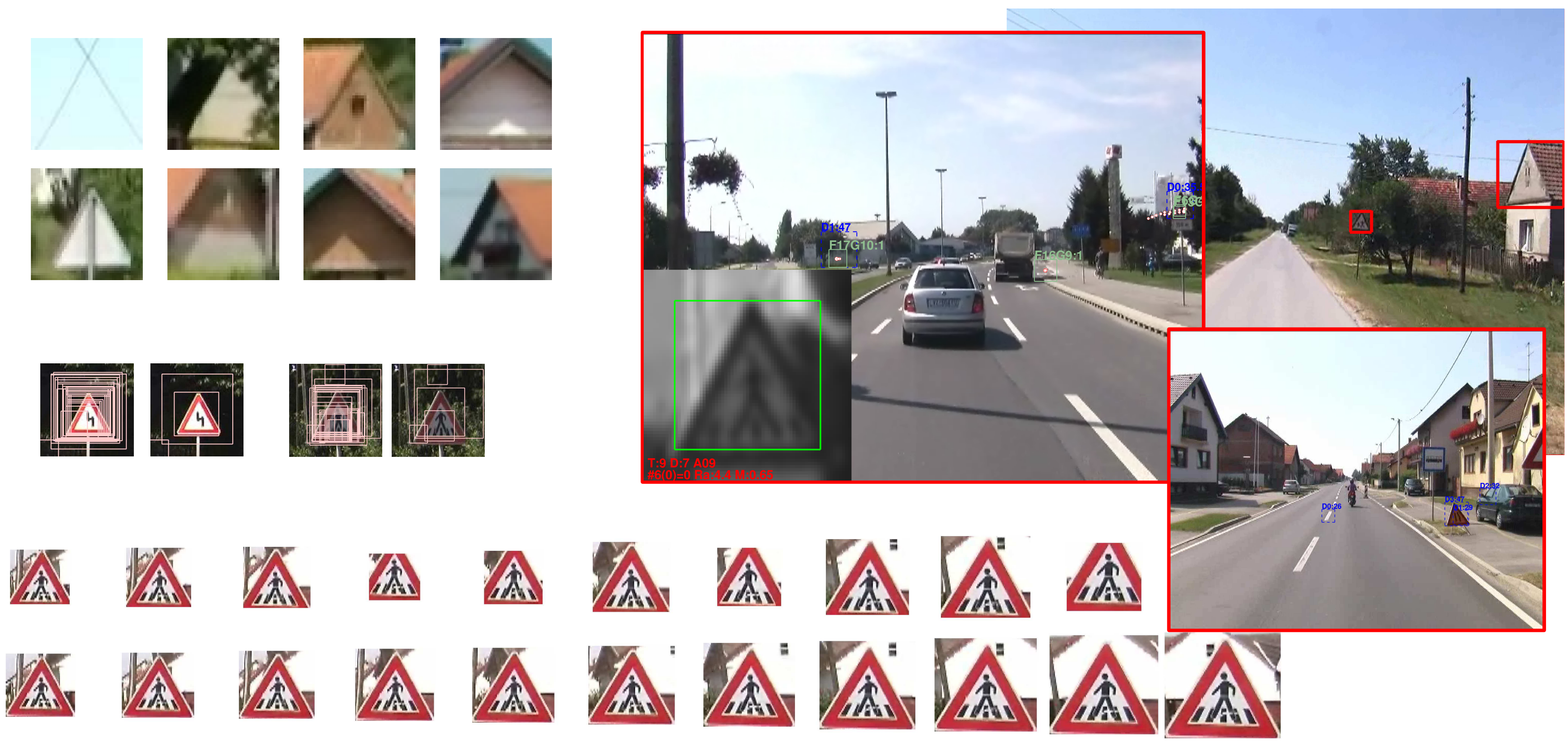


mapping: create the traffic inventory from a recorded geo-referenced video



Introduction: challenges

- low precision (false positives)
- multiple responses
- localization inaccuracy
- lateral displacement
- distance along the optical axis
- non-standard orientation



INTRODUCTION: OPEN QUESTIONS

- multi-class detection of ideogram-based signs
- a principled approach to deal with layout variability
- detecting foreground motion

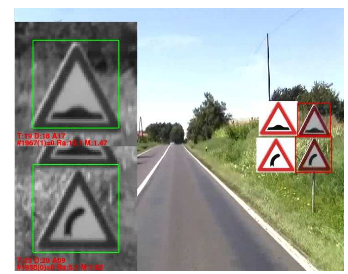
Data: assumptions

We consider SDTV video acquired from the driver's perspective along the Croatian local roads (720×576 pixels, HFOV$=48^\circ$)

Data: assumptions

We consider SDTV video acquired from the driver's perspective along the Croatian local roads (720×576 pixels, HFOV=48°)

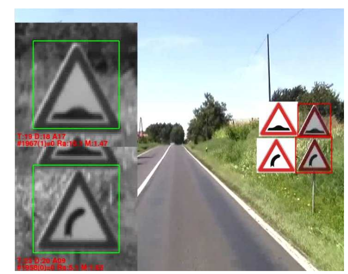
Typically, signs leave the field of view when they are about 80×80 pixels large (may be smaller due to lateral displacement)



Data: assumptions

We consider SDTV video acquired from the driver's perspective along the Croatian local roads (720×576 pixels, HFOV=48°)

Typically, signs leave the field of view when they are about 80×80 pixels large (may be smaller due to lateral displacement)

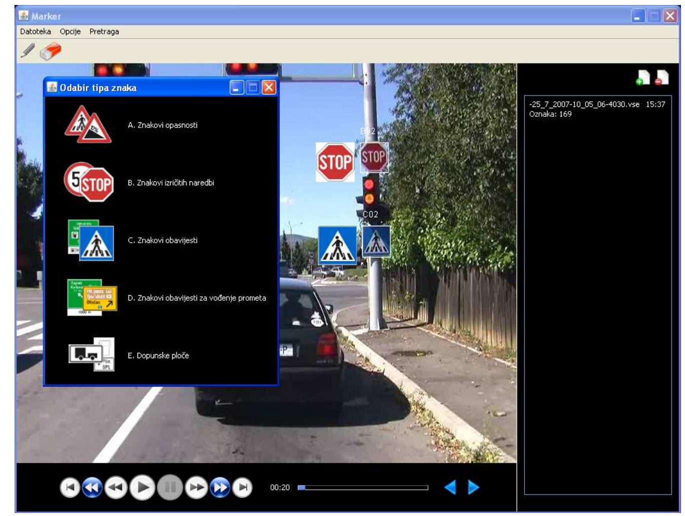


One has to deal with noisy pixels, motion blur and **unreliable colours**

DATA: ANNOTATION

Large sample collections required for proper test and training

We developed a custom software tool (Marker) to collect samples from video



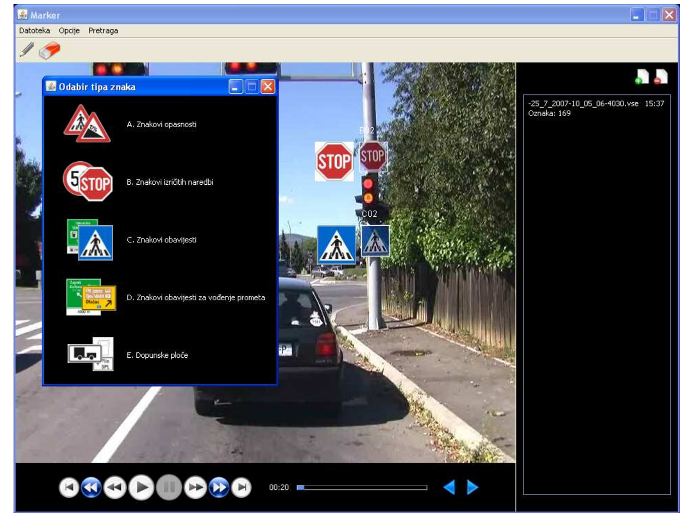
We systematically annotated many hours of production video provided by partners
(all kinds of traffic signs were annotated)

Data: annotation

Large sample collections required for proper test and training

We developed a custom software tool (Marker) to collect samples from video

We systematically annotated many hours of production video provided by partners
(all kinds of traffic signs were annotated)



Each physical sign is annotated four times as follows:

We collected about 7500 annotations of different sign classes.

DATA: FOCUS

We focus on the class of danger warning signs since:

- most frequent: 3000 of 7500 annotations total (almost 50%)
- well standardized according to the Vienna Convention (1968)
- research results likely relevant for other ideogram-based signs

Data: focus

We focus on the class of danger warning signs since:

- most frequent: 3000 of 7500 annotations total (almost 50%)
- well standardized according to the Vienna Convention (1968)
- research results likely relevant for other ideogram-based signs

This leaves out only the direction signs, some signs from the information class and additional panels.

DATA: DATASETS

We organize the 3000 annotated samples of danger warning signs into two datasets:

- T2009: 2000 signs acquired with interlaced camera
- T2010: 1000 signs acquired with a progressive camera
- we use T2009 for training (left), T2010 for evaluation (right)

Data: datasets

We organize the 3000 annotated samples of danger warning signs into two datasets:

- T2009: 2000 signs acquired with interlaced camera
- T2010: 1000 signs acquired with a progressive camera
- we use T2009 for training (left), T2010 for evaluation (right)

Both the datasets and our annotation program can be freely downloaded from the web site of our research project:

- project home: `http://www.zemris.fer.hr/~ssegvic/mastif/index_en.shtml`
- datasets: `http://www.zemris.fer.hr/~ssegvic/mastif/datasets.shtml`
- marker: `http://www.zemris.fer.hr/~ssegvic/mastif/marker/marker.zip`

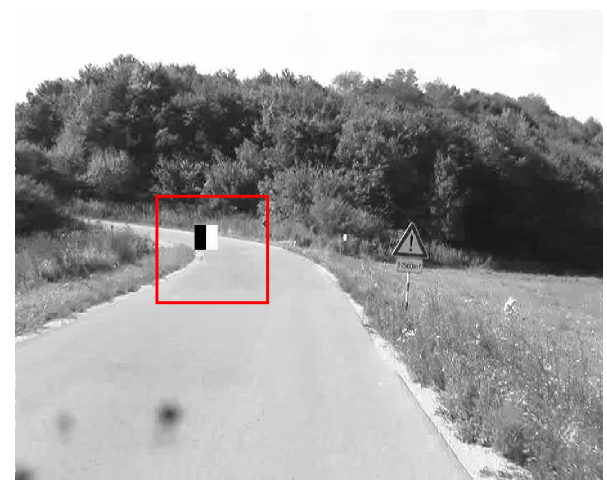
Baseline detection: alternatives

Approaches based on detecting primitives such as colour and geometry resulted in insufficient detection and poor precision:

- colour-based detection with hardwired thresholds over HSI
- Hough transform approach for circular signs
- radial symmetry for triangular signs

Much better results achieved when looking at pixels directly:

- sliding window approach: binary classification at all image positions and scales
- advantage: work directly with sensed data (focus on grey-scale appearance)



- liabilities: complexity (10^6 queries/image), large training datasets

Baseline detection: BHC pros

Boosted Haar cascades: a great approach to detect objects in images

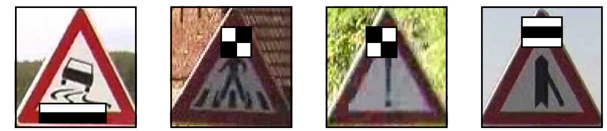
Baseline detection: BHC pros

Boosted Haar cascades: a great approach to detect objects in images

- **Haar classifier**: Haar feature + threshold + polarity 

Baseline detection: BHC pros

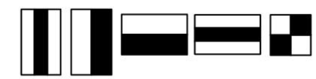
Boosted Haar cascades: a great approach to detect objects in images

- **Haar classifier**: Haar feature + threshold + polarity 
- **boosted classifier**: an ensemble of simple Haar classifiers 

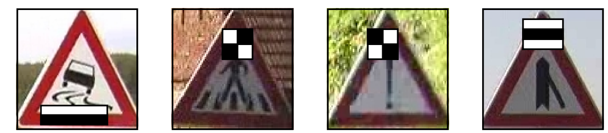
Baseline detection: BHC pros

Boosted Haar cascades: a great approach to detect objects in images

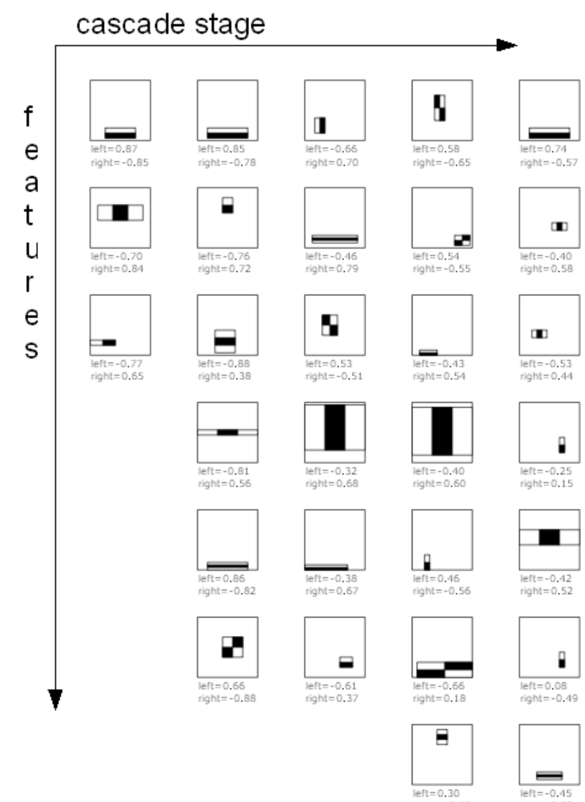
- **Haar classifier**: Haar feature + threshold + polarity



- **boosted classifier**: an ensemble of simple Haar classifiers



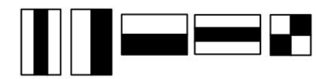
- the **cascade** consists of boosted classifiers with increasing complexity (most queries will be negative!)
- complexity is tuned by training each stage on false positives of its predecessors!



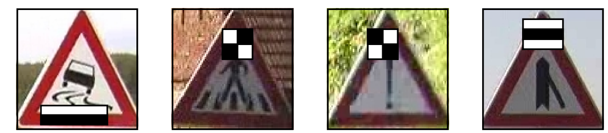
Baseline detection: BHC pros

Boosted Haar cascades: a great approach to detect objects in images

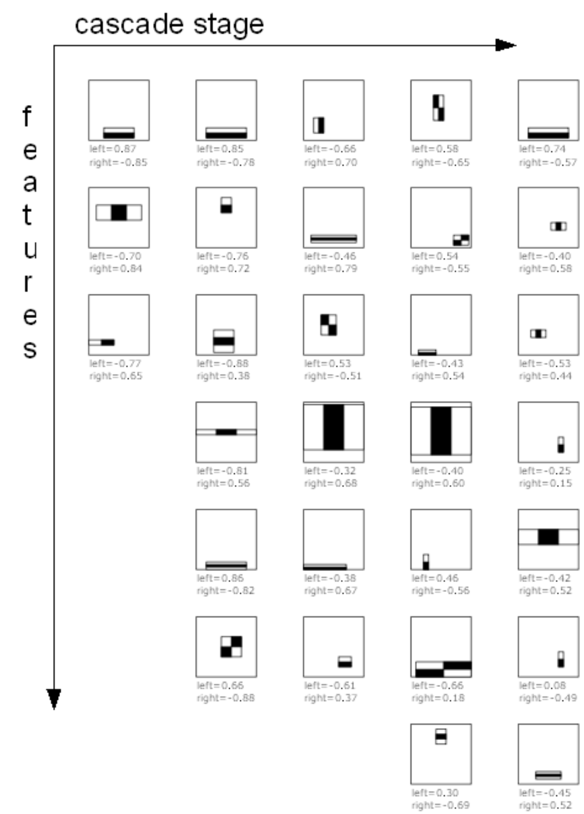
- **Haar classifier**: Haar feature + threshold + polarity



- **boosted classifier**: an ensemble of simple Haar classifiers



- the **cascade** consists of boosted classifiers with increasing complexity (most queries will be negative!)



- complexity is tuned by training each stage on false positives of its predecessors!
- excellent ratio of performance vs computational burden (720×576 images, quad CPU: 50 ms)
- encouraging **recall**: over 95% signs detected [itsc11]

Baseline detection: BHC cons

Although very good, boosted Haar cascades do not provide enough performance for automated operation:

- strong dependence on sign size
  - colour may help only with large signs [bonaci11cvww]

Baseline detection: BHC cons

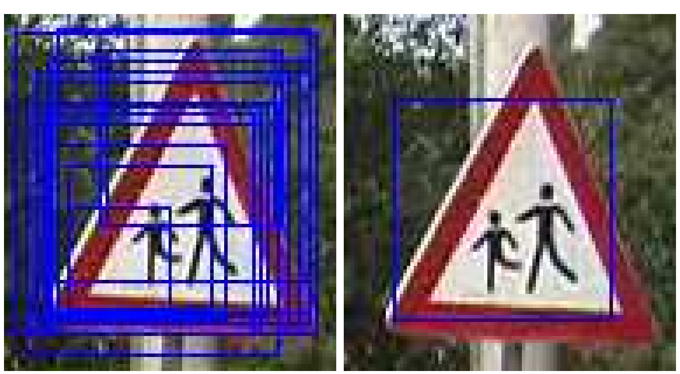
Although very good, boosted Haar cascades do not provide enough performance for automated operation:

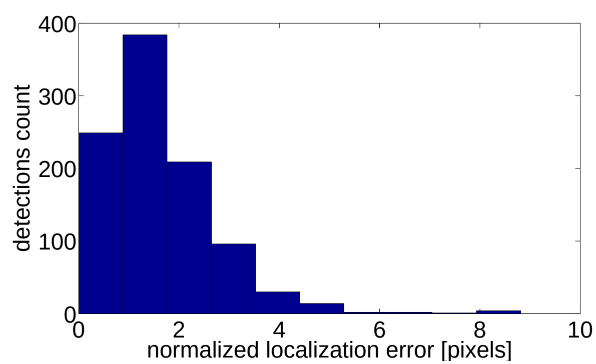
- strong dependence on sign size
  - colour may help only with large signs [bonaci11cvww]
- unsatisfactory precision, 50% or lower
  - BHCs poor at generalizing over unseen negatives

Baseline detection: BHC cons

Although very good, boosted Haar cascades do not provide enough performance for automated operation:

- strong dependence on sign size
  - colour may help only with large signs [bonaci11cvww]
- unsatisfactory precision, 50% or lower
  - BHCs poor at generalizing over unseen negatives
- localization accuracy leaves to desire:
  - we care because bad localization hurts recognition [itsc10]!





Method: approach

Cascading classifiers of increasing complexity works great.

Method: approach

Cascading classifiers of increasing complexity works great.

The **proposed approach** follows the same track:

- configure BHC for high recall (skip heuristic grouping!)
- devise additional techniques to improve precision and localization

Method: approach

Cascading classifiers of increasing complexity works great.

The **proposed approach** follows the same track:

- configure BHC for high recall (skip heuristic grouping!)
- devise additional techniques to improve precision and localization

These additional techniques can be computationally expensive without hurting overall performance!

- our BHC-s (2000 training samples, 95% recall) typically let by less than 10 false positives per image!

Method: approach

Cascading classifiers of increasing complexity works great.

The **proposed approach** follows the same track:

- configure BHC for high recall (skip heuristic grouping!)
- devise additional techniques to improve precision and localization

These additional techniques can be computationally expensive without hurting overall performance!

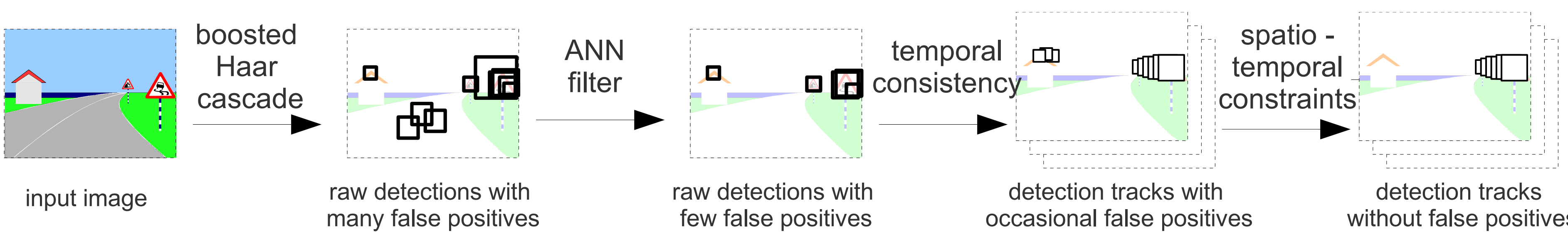
- our BHC-s (2000 training samples, 95% recall) typically let by less than 10 false positives per image!

The concept of heterogeneous classification cascades can be further applied at the level of temporal detection sequences in video!

Method: the big figure

The devised detection pipeline:

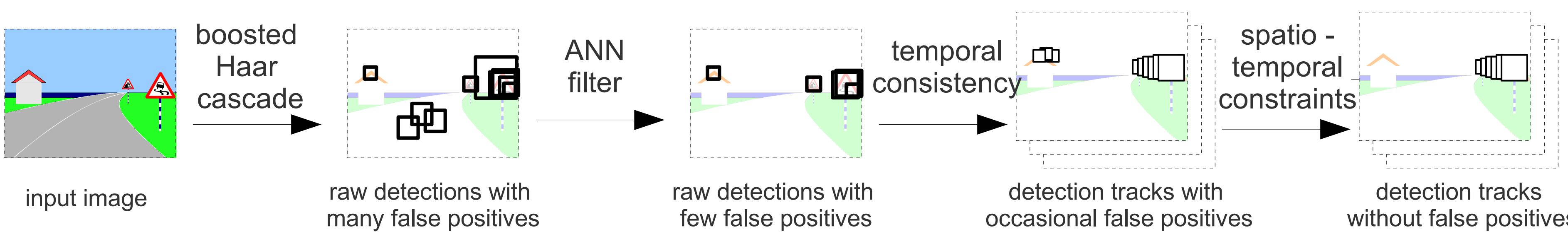
- baseline detection by boosted Haar cascades
- introduce a strong classifier in the additional cascade stage to improve *precision*
- enforce temporal consistency by differential tracking to improve *localization accuracy* and further improve *precision*
- enforce learned contextual constraints to further improve *precision*



Method: the big figure

The devised detection pipeline:

- baseline detection by boosted Haar cascades
- introduce a strong classifier in the additional cascade stage to improve *precision*
- enforce temporal consistency by differential tracking to improve *localization accuracy* and further improve *precision*
- enforce learned contextual constraints to further improve *precision*

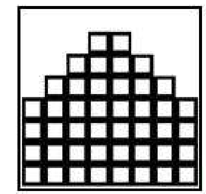


The last two stages operate on detection tracks: temporal sequences of traffic sign position, scale and appearance

Method: additional strong classifier

A heterogeneous cascade for object detection in images [bonaci11cvww]:

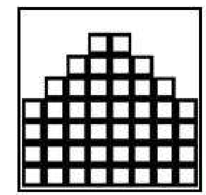
- use boosted Haar cascade for fast rejection of easy negatives
- use a strong classifier to decide about the **hard** cases
  - suitable ANN applied to a HOG descriptor
  - similar results achieved by SVM+HOG



METHOD: ADDITIONAL STRONG CLASSIFIER

A heterogeneous cascade for object detection in images [bonaci11cvww]:

- use boosted Haar cascade for fast rejection of easy negatives
- use a strong classifier to decide about the **hard** cases
  - suitable ANN applied to a HOG descriptor
  - similar results achieved by SVM+HOG



How do we combine the BHC and ANN+HOG?

- train a BHC for max recall and reasonable precision on T2009
- train ANN+HOG on BHC false positives collected on T2009
- perform detection by applying ANN+HOG to the BHC survivors
- **important**: the above must be performed **before** the grouping step

Method: additional strong classifier (2)

The results:

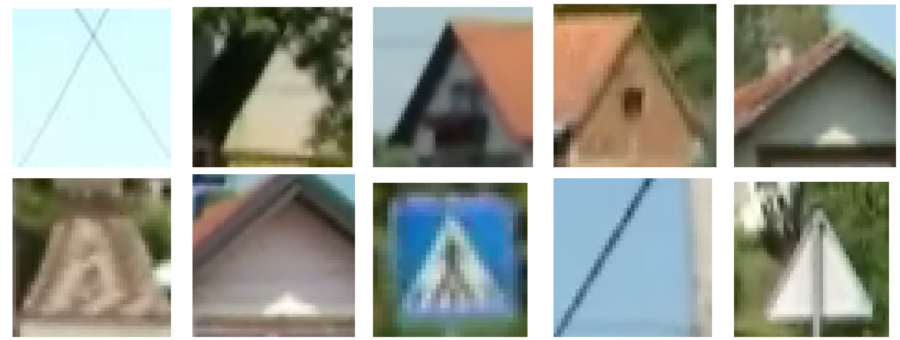
- precision: 57% $\longrightarrow$ 89%!
- **recall**: only slightly worse!
- **localization**: slightly better!

Method: additional strong classifier (2)

The results:

- precision: 57% → 89%!
- **recall**: only slightly worse!
- **localization**: slightly better!

The surviving false positives:

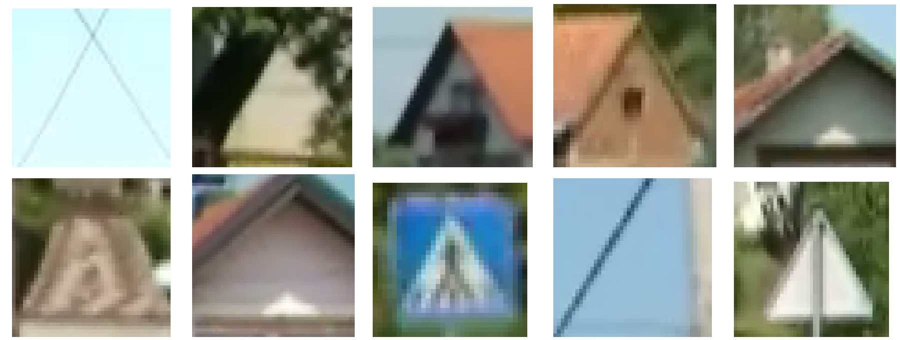


METHOD: ADDITIONAL STRONG CLASSIFIER (2)

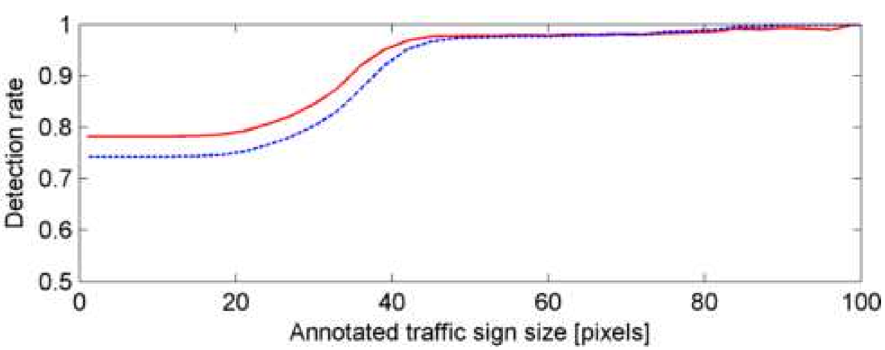
The results:

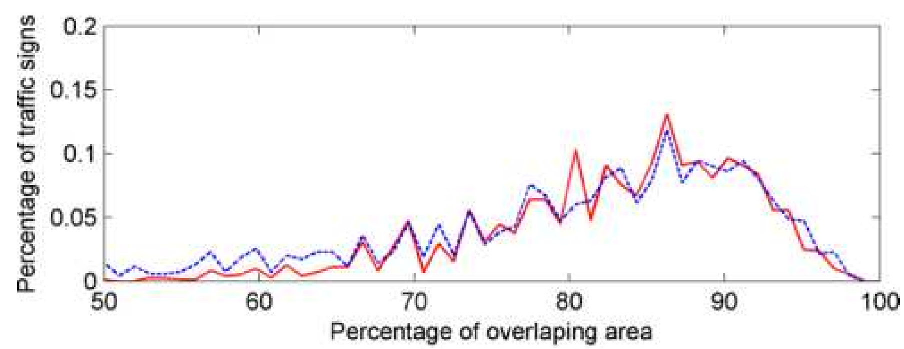
- precision: 57% $\rightarrow$ 89%!
- **recall**: only slightly worse!
- **localization**: slightly better!

The surviving false positives:



Detection rate (left) and localization accuracy (right), BHC (red) and BHC || ANN+HOG (blue), depending on the sign size:





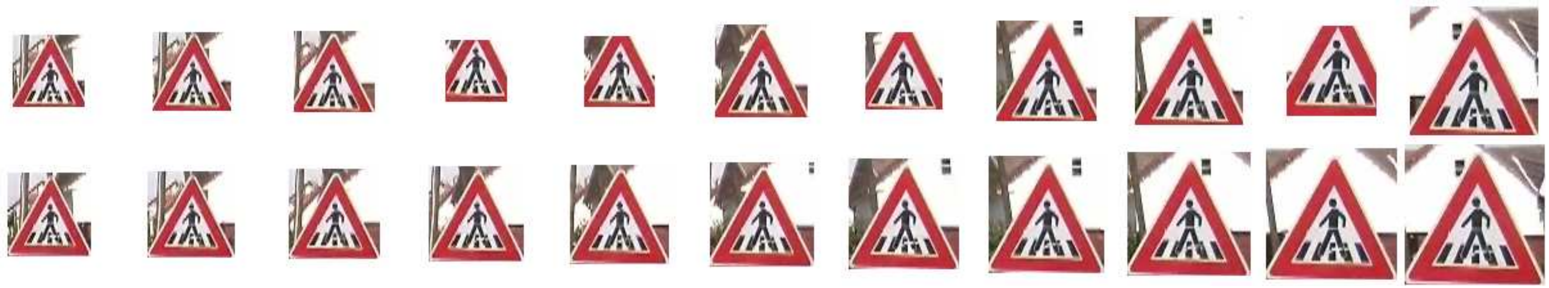
METHOD: ADDITIONAL STRONG CLASSIFIER (3)

Some results (blue: BHC, red: ANN+HOG):

Method: consistency

Idea: require that detection sequences be temporally consistent [mva11]

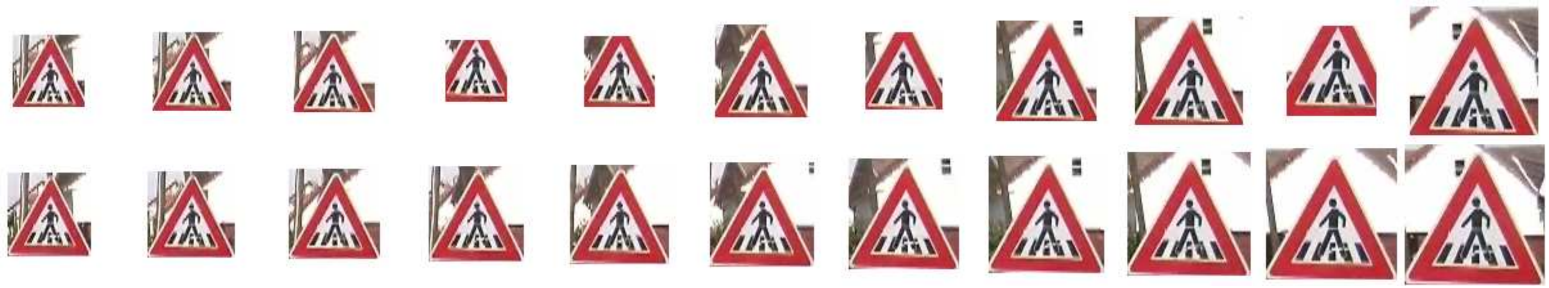
- top: raw detection chain, bottom: the desired detection track



Method: consistency

Idea: require that detection sequences be temporally consistent [mva11]

- top: raw detection chain, bottom: the desired detection track



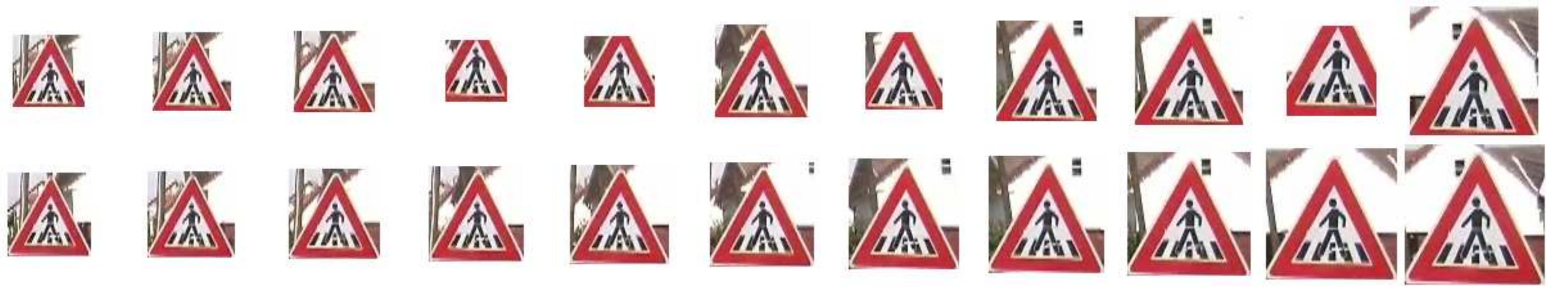
Approach:

- track many detection hypotheses along the sequence
- pick the one which receives most detections!

Method: consistency

Idea: require that detection sequences be temporally consistent [mva11]

- top: raw detection chain, bottom: the desired detection track



Approach:

- track many detection hypotheses along the sequence
- pick the one which receives most detections!

Benefits in comparison to detection chaining:

- reject false positives which are i) temporally inconsistent or ii) large
- better localization due to i) lack of grouping, and ii) integrating evidence from many frames

Method: consistency (2)

Implementation details:

- seed a new detection track hypothesis in the interior of each detection displaced from all active hypotheses

Method: consistency (2)

Implementation details:

- seed a new detection track hypothesis in the interior of each detection displaced from all active hypotheses
- track all hypotheses in parallel by combining the detector and the tracker (somewhat in the spirit of particle filter)

Method: consistency (2)

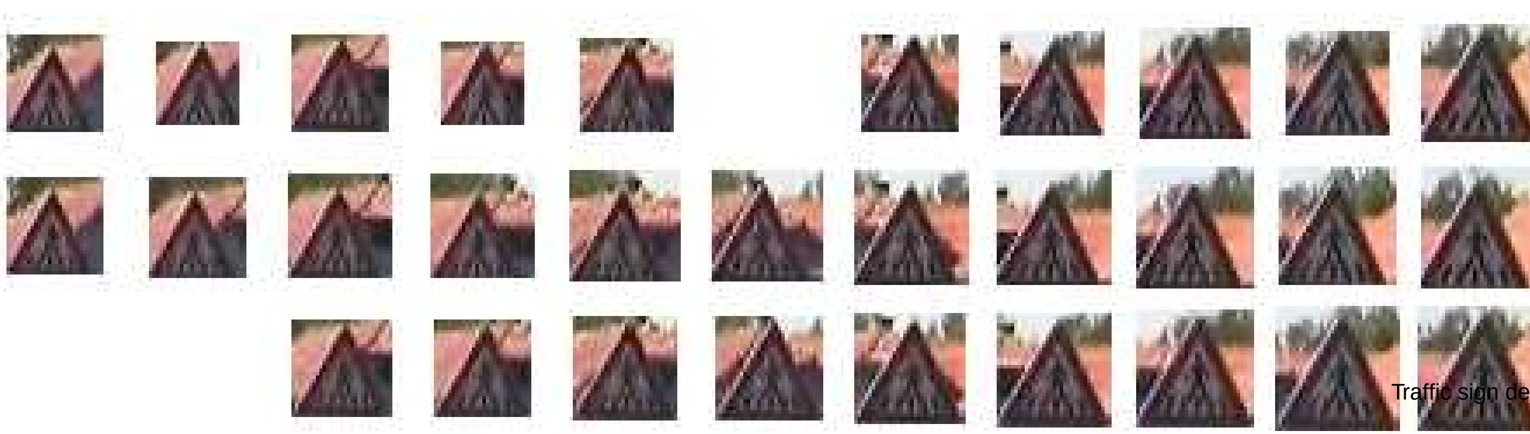
Implementation details:

- seed a new detection track hypothesis in the interior of each detection displaced from all active hypotheses
- track all hypotheses in parallel by combining the detector and the tracker (somewhat in the spirit of particle filter)
- group overlapping hypotheses into clusters corresponding to distinct physical signs

METHOD: CONSISTENCY (2)

Implementation details:

- seed a new detection track hypothesis in the interior of each detection displaced from all active hypotheses
- track all hypotheses in parallel by combining the detector and the tracker (somewhat in the spirit of particle filter)
- group overlapping hypotheses into clusters corresponding to distinct physical signs
- when all hypotheses of a cluster are lost, pick the hypothesis with most evidence from raw detections



Method: consistency (3)

Results:

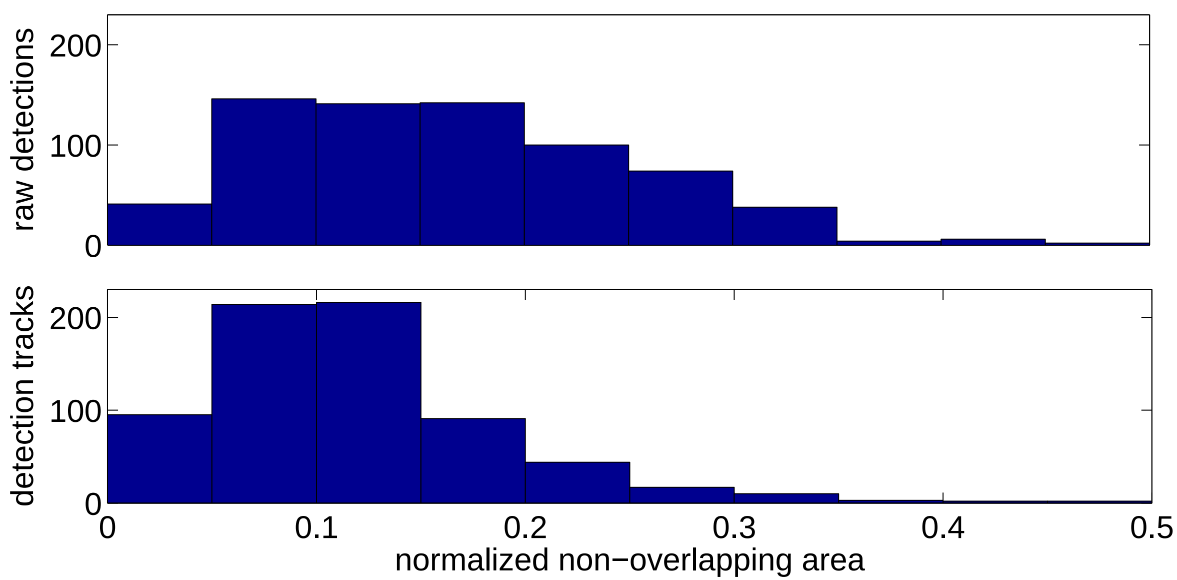
- near 100% recall on the system level
- 2 false positives in 11000 traffic images
  (vs 14 with a criterion based on detection chains)
- measurable improvement in localization accuracy

Method: consistency (3)

Results:

- near 100% recall on the system level
- 2 false positives in 11000 traffic images
  (vs 14 with a criterion based on detection chains)
- measurable improvement in localization accuracy

Raw detection responses (top) vs detection tracks (bottom):



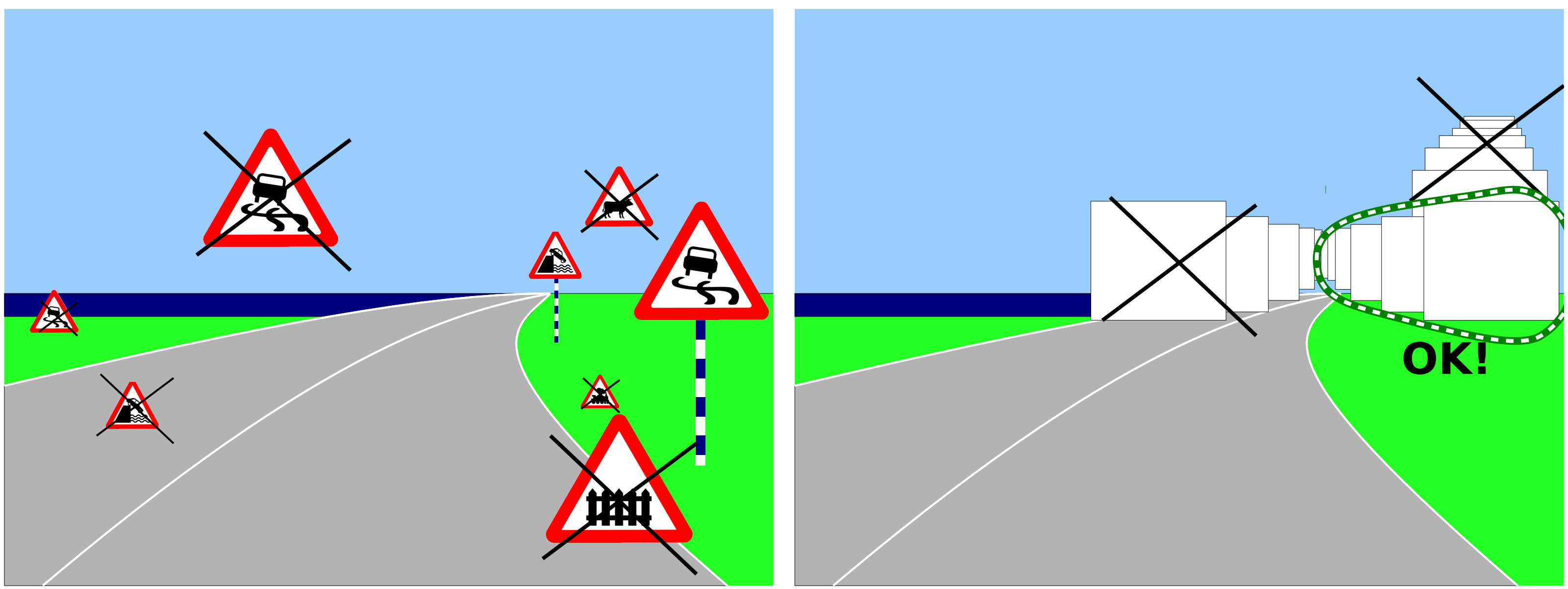
Method: consistency (4)

Some hard cases:

Method: spatio-temporal constrains

Focus on spatio-temporal properties of traffic sign occurences:

- at which image locations and scales the signs typically occur?
- which typical trajectories do the signs follow?
- learn a discriminative model for classifying detection tracks into signs and not-signs



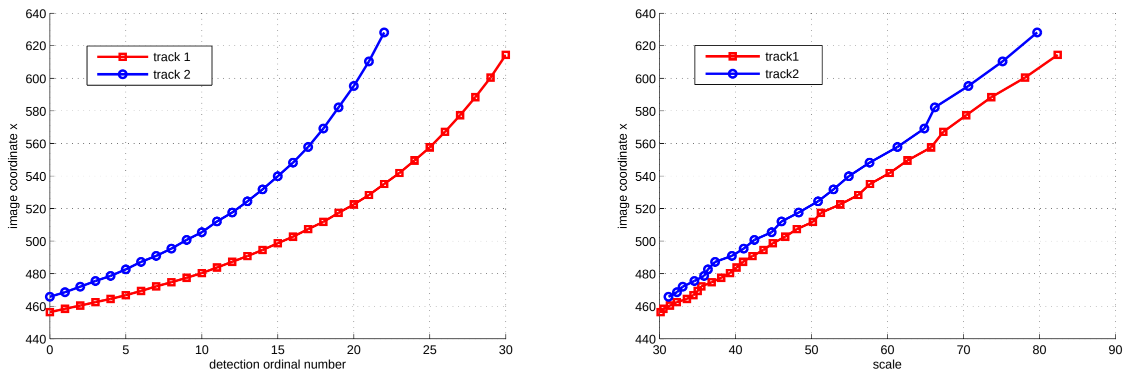
METHOD: SPATIO-TEMPORAL CONSTRAINS (2)

Camera type and placement do not change $\Rightarrow$ can reason in pixels!

Method: spatio-temporal constrains (2)

Camera type and placement do not change ⇒ can reason in pixels!

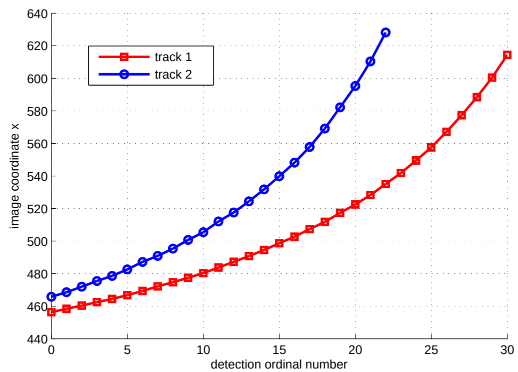
Vehicle speed does change ⇒ look at sequences of x/scale and y/scale!

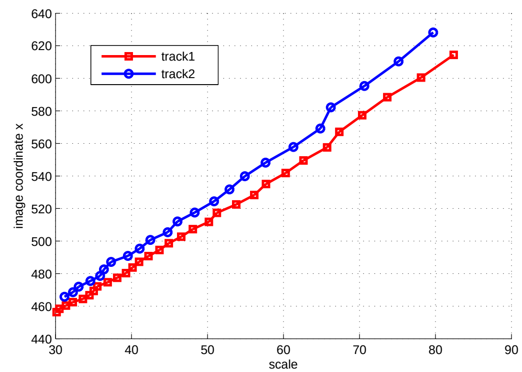


Method: spatio-temporal constrains (2)

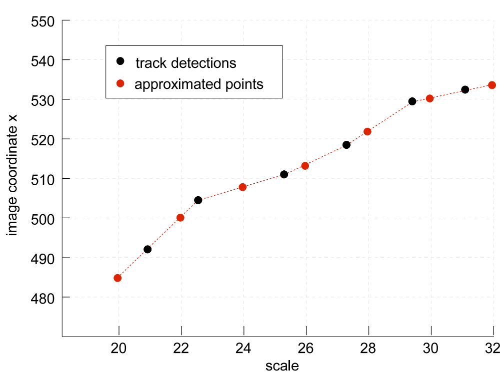
Camera type and placement do not change ⇒ can reason in pixels!

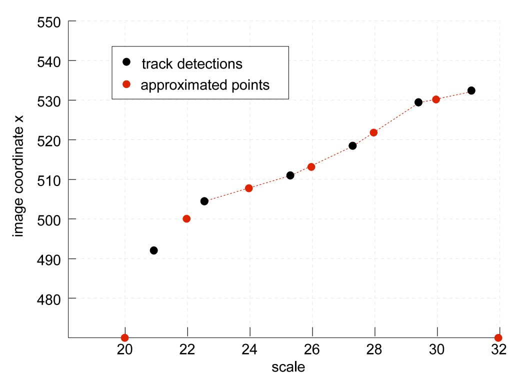
Vehicle speed does change ⇒ look at sequences of x/scale and y/scale!





Not all signs are visible at all scales ⇒ must either extrapolate or impute unknown data points!





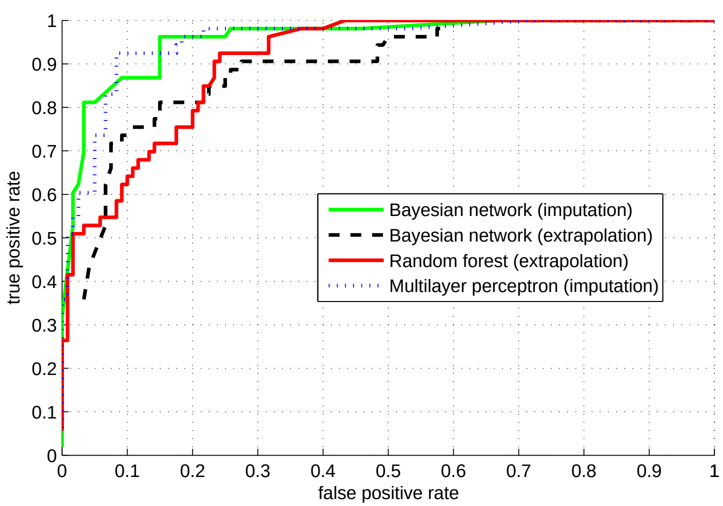
Method: spatio-temporal constrains (3)

We tested the concept before we developed the strong classifier in the additional cascade stage

Method: spatio-temporal constrains (3)

We tested the concept before we developed the strong classifier in the additional cascade stage

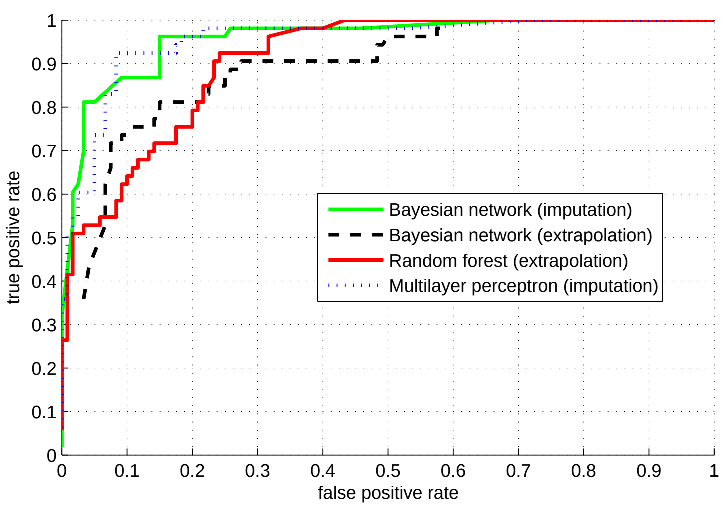
Best recognition achieved with Bayesian networks and imputation



Method: spatio-temporal constrains (3)

We tested the concept before we developed the strong classifier in the additional cascade stage

Best recognition achieved with Bayesian networks and imputation



The best classifier managed to discard 82% false positives while retaining 98% recall

Towards multi-class detection: overview

Why would we like to have one multi-class detector instead of n single-class detectors?

Towards multi-class detection: overview

Why would we like to have one multi-class detector instead of n single-class detectors?

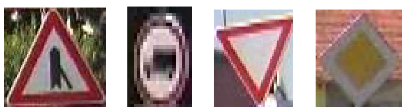
Because for ideogram-based traffic signs $n > 20$!

    ...

Towards multi-class detection: overview

Why would we like to have one multi-class detector instead of n single-class detectors?

Because for ideogram-based traffic signs $n > 20$!

 ...

How about parallelization?

- MIMD (multicore): linear detection speedup on a quad core CPU
  - however, affordable many-cores are not coming anytime soon
- SIMD (GPU): not suitable for implementing cascades [ghorayeb06accv]

Towards multi-class detection: overview

Why would we like to have one multi-class detector instead of n single-class detectors?

Because for ideogram-based traffic signs $n > 20$!

    ...

How about parallelization?

- MIMD (multicore): linear detection speedup on a quad core CPU
  - however, affordable many-cores are not coming anytime soon
- SIMD (GPU): not suitable for implementing cascades [ghorayeb06accv]

To conclude, advances towards logarithmic increase of complexity with respect to n would be — very interesting!

Towards multi-class detection: individual detectors

class	# training	# evaluation	recall	false alarms/image
	2150	886	96.2%	4.4
	645	377	100%	9.7
	106	8	87.5%	12.1
	337	49	98.0%	12.9

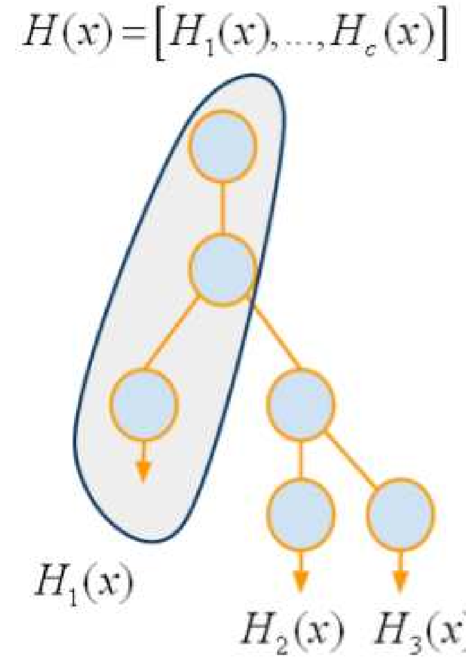
For homogeneous classes (last two rows), fairly good results can be obtained even with few training samples!

Towards multi-class detection: CBT

Cluster boosted trees [wu07iccv]:
a classification approach based on feature sharing

Major advantage with respect to JointBoost:
suitable for detection in a sliding window

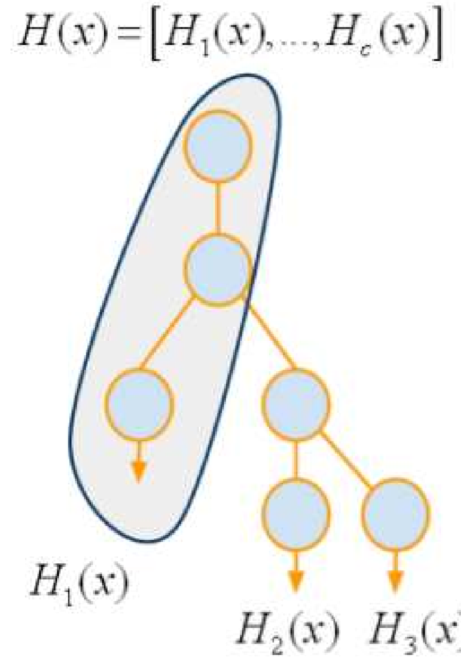
- the classification gradually focuses, no need to calculate all features to evaluate a query!



Towards multi-class detection: CBT

Cluster boosted trees [wu07iccv]:

a classification approach based on feature sharing



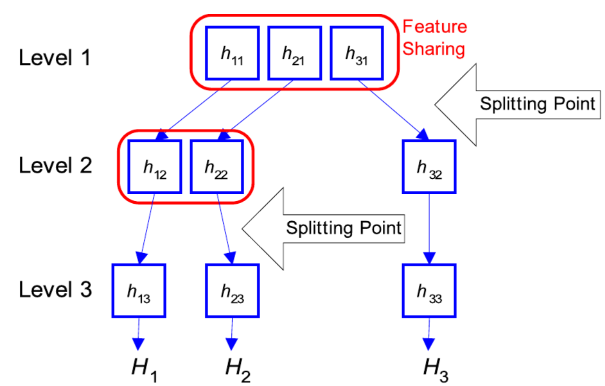
Major advantage with respect to JointBoost:

suitable for detection in a sliding window

- the classification gradually focuses, no need to calculate all features to evaluate a query!

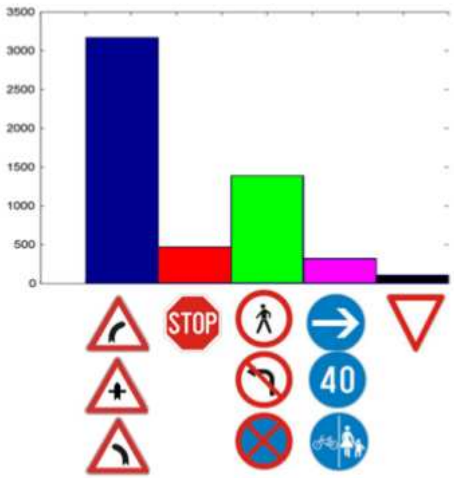
The training proceeds like in usual boosting except that:

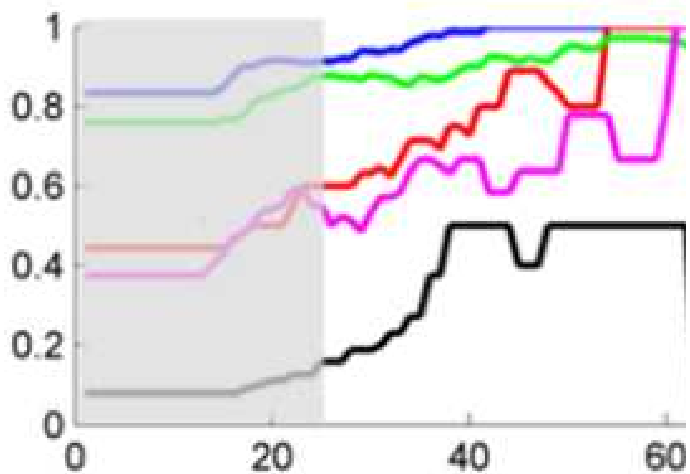
- the tree is split whenever a newly added node has low discriminative power
- after the tree is constructed, the thresholds are separately retrained for each leaf class

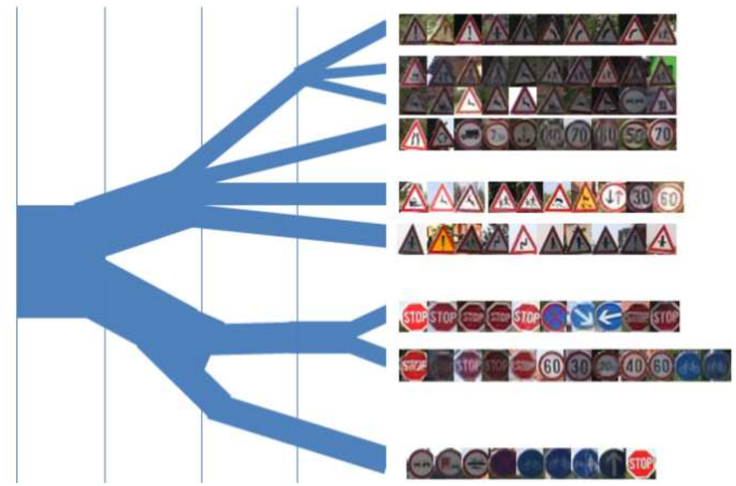


TOWARDS MULTI-CLASS DETECTION: CBT (2)

The achieved performance (Haar classifiers) and the resulting tree:

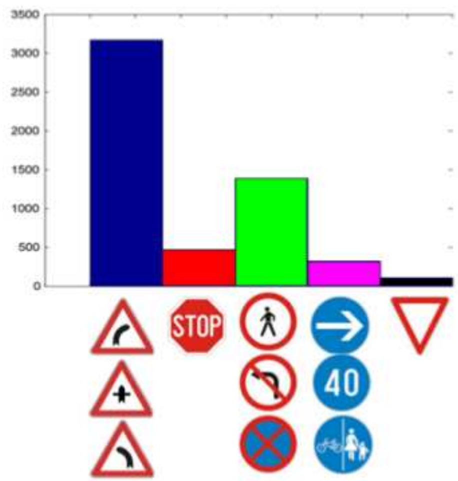


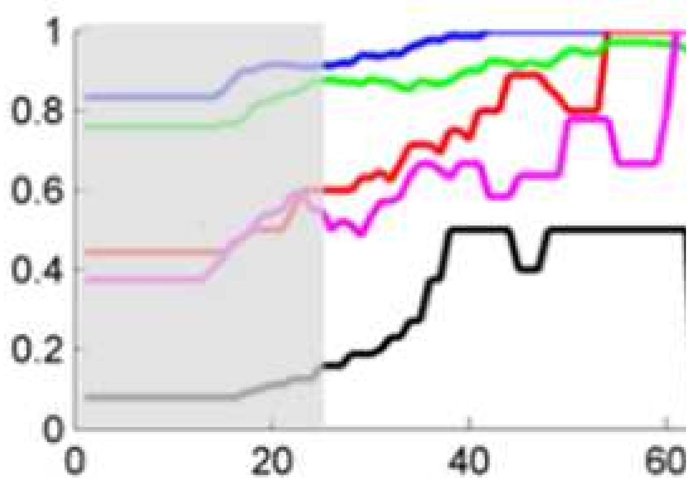


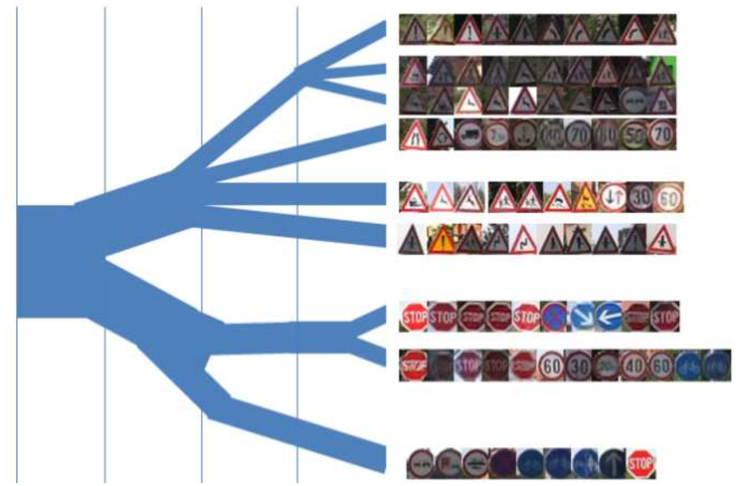


TOWARDS MULTI-CLASS DETECTION: CBT (2)

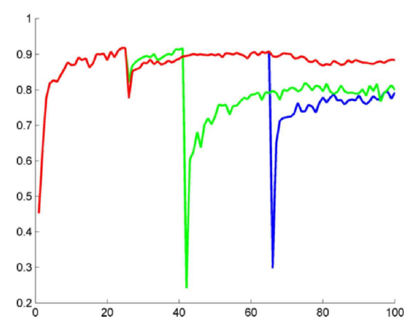
The achieved performance (Haar classifiers) and the resulting tree:



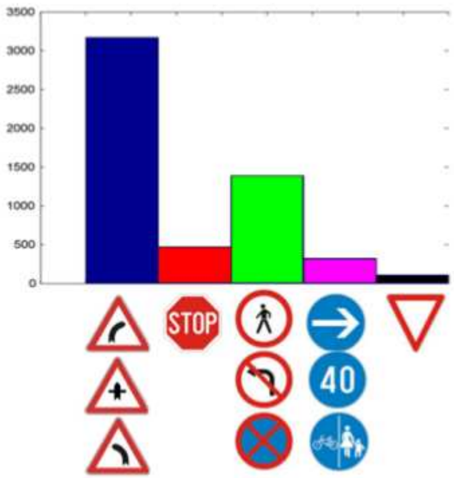


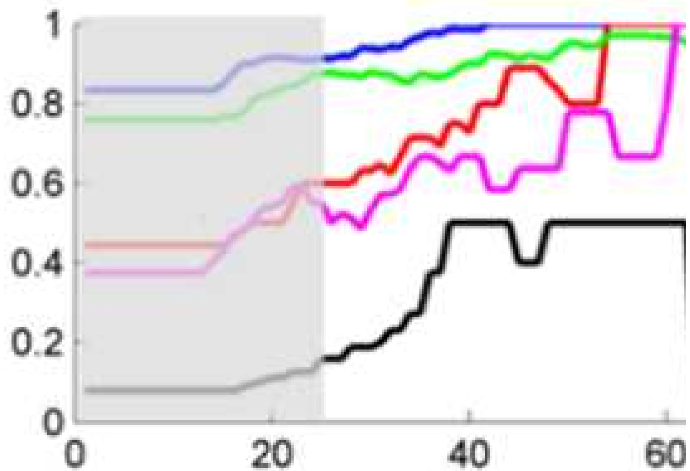


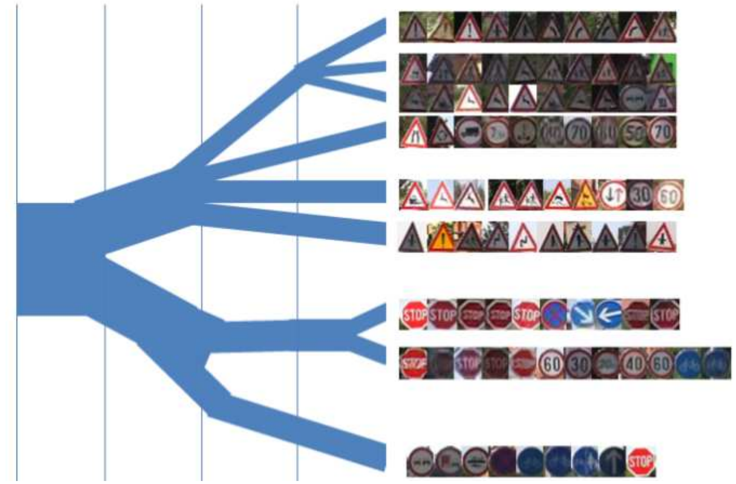
Branch splitting occurs when a newly added feature is not discriminative (test is based on Bhattacharya distance)



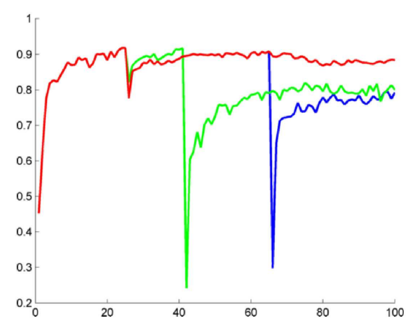
The achieved performance (Haar classifiers) and the resulting tree:







Branch splitting occurs when a newly added feature is not discriminative (test is based on Bhattacharya distance)

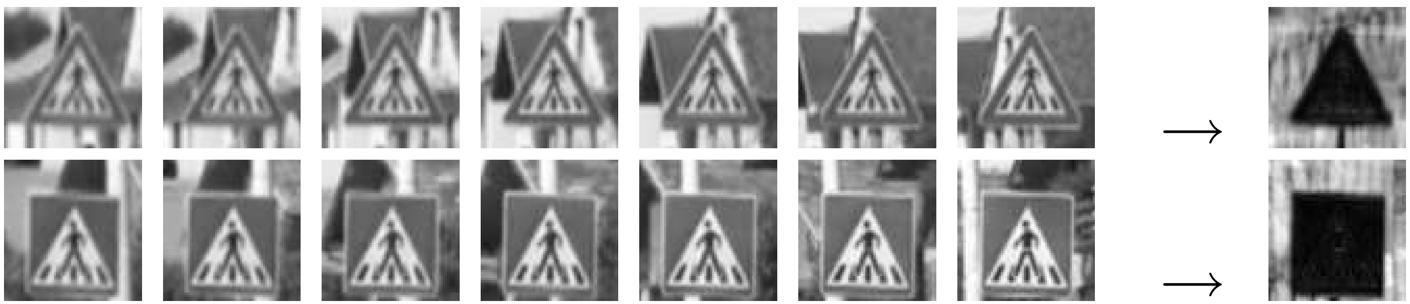


There is a big performance gap between shared and dedicated features!

- 50% vs 90% for the yield sign
- a possible way to deal with that: introduce more complex features in advanced stages

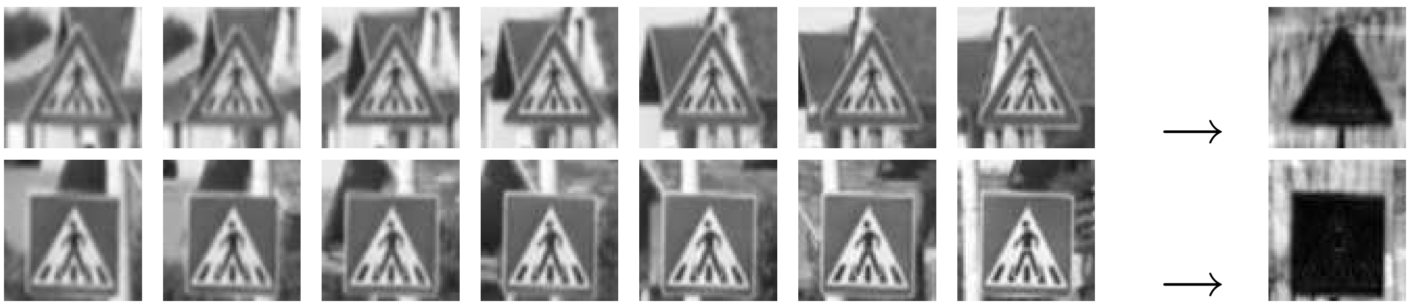
Towards multi-class detection: generic detection

By looking at the pixel variance one can recover the shape of the tracked object [mva11]:



Towards multi-class detection: generic detection

By looking at the pixel variance one can recover the shape of the tracked object [mva11]:



We curently research ways to employ this concept for bottom-up detection of occluding shapes [brkic11scia]

- great potential for detecting all kinds of table-like objects!
- precondition: successful tracking of features at signs

Conclusion

Single-class detection of ideogram-based traffic signs:

- baseline detection (BHC) achieves about 95% recall with mean relative displacement of 17% and about 1 false positive per image
  - if only large signs are considered, the recall approaches 100%
- additional filter (ANN+HOG) reduces the false positive incidence to about 1 in 9 images, while retaining recall
- a criterion based on detection chains reduces false positives to about 1 in 700 images
- temporal consistency reduces false positives to about 1 in 50000, and improves the mean relative displacement to 12%
- spatio-temporal constraints show potential for resolving the remaining false positives

Future work

Bridging the gap between multi-class detection with shared-features and dedicated per-class detectors

Future work

Bridging the gap between multi-class detection with shared-features and dedicated per-class detectors

Generic detection of table-like objects

Future work

Bridging the gap between multi-class detection with shared-features and dedicated per-class detectors

Generic detection of table-like objects

Detecting and recognizing direction tables regardless of colour

Future work

Bridging the gap between multi-class detection with shared-features and dedicated per-class detectors

Generic detection of table-like objects

Detecting and recognizing direction tables regardless of colour

Detecting and recognizing lane configuration signs

Thank you for your attention!

This work has been jointly performed by
Karla Brkić, Zoran Kalafatić, Axel Pinz and the presenter.

Parts of this work have been performed by our undergraduate students
Igor Bonači, Ivan Kovaček and Ivan Kusalić.

We are grateful for the support by Croatian Science Foundation,
Institute of Traffic and Communications, and Graz University of Technology.

References

Siniša Šegvić, Karla Brkić, Zoran Kalafatic, Axel Pinz. Exploiting temporal and spatial constraints in traffic sign detection from a moving vehicle. Machine Vision and Applications. Accepted for publication.

Igor Bonači, Ivan Kusalić, Ivan Kovaček, Zoran Kalafatić, Siniša Šegvić. Addressing false alarms and localization inaccuracy in traffic sign detection and recognition. CVWW, Mitterberg, Austria, January 2011.

Karla Brkić, Axel Pinz, Siniša Šegvić, Zoran Kalafatić. Histogram-Based Description of Local State-Time Appearance. SCIA, Ystad Saltsjöbad, Sweden May 2011.

Ivan Sikirić, Ante Majić, Siniša Šegvić. Using GPS positioning to recover a comprehensive road appearance mosaic. MIPRO, Opatija, Croatia, May 2011.

Petar Palašek, Petra Bosilj, Siniša Šegvić. Detecting and recognizing centerlines as parabolic sections of the steerable filter response. MIPRO, Opatija, Croatia, May 2011.

Siniša Šegvić, Karla Brkić, Zoran Kalafatić, Vladimir Stanisavljević, Marko Ševrović, Damir Budimir and Ivan Dadić. A computer vision assisted geoinformation inventory for traffic infrastructure. ITSC 2010, Madeira, Portugal, September 2010.