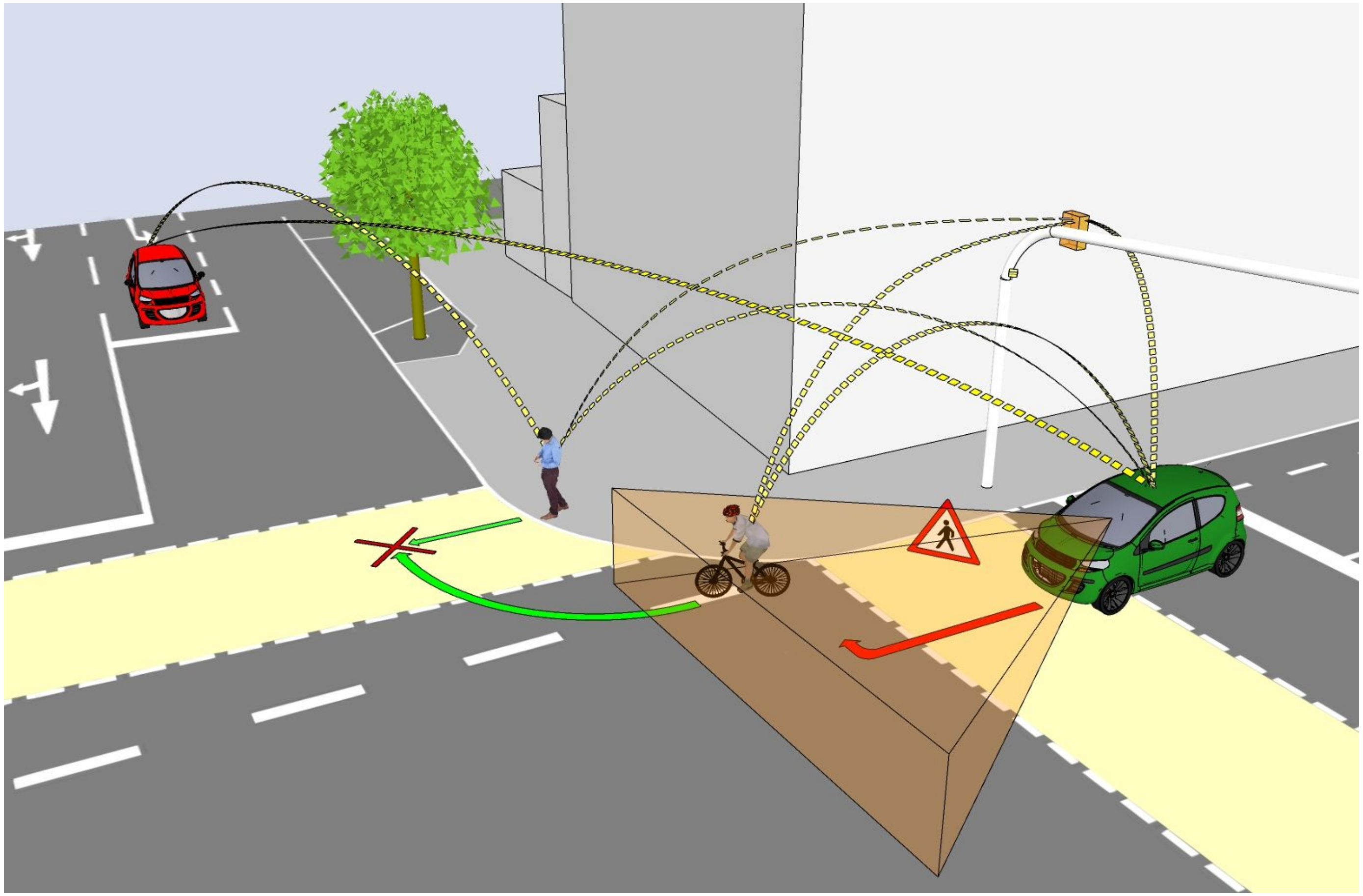
Image Sequence Based Cyclist Action Recognition Using Multi-Stream 3D Convolution

Stefan Zernetsch, Steven Schreck, Viktor Kress, Konrad Doll, and Bernhard Sick

ZD.B ZENTRUM DIGITALISIERUNG. BAYERN
DFG Deutsche Forschungsgemeinschaft

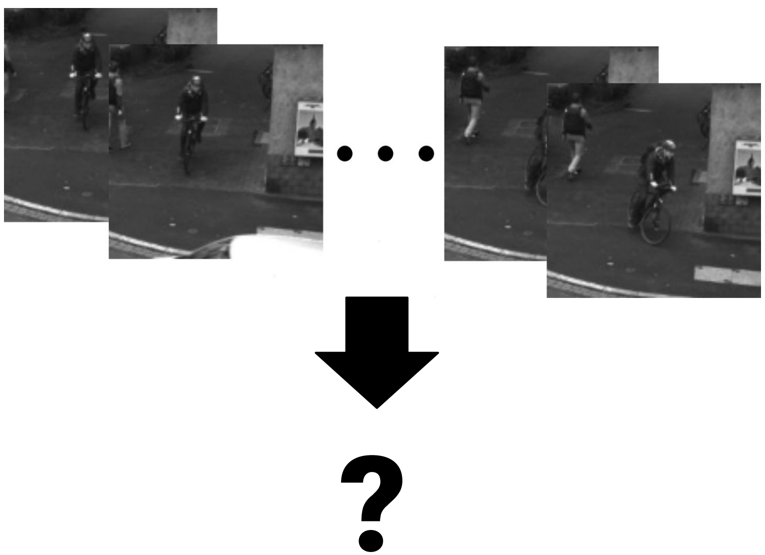
TH Aschaffenburg
university of applied sciences

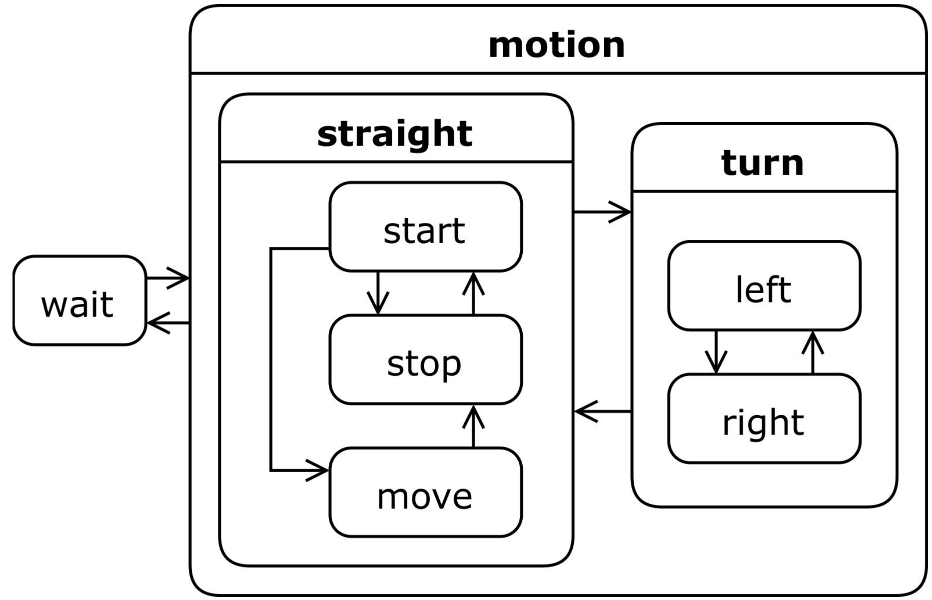
Motivation



Cooperative intention detection of vulnerable road users in urban areas as basis for automated driving.

Goals



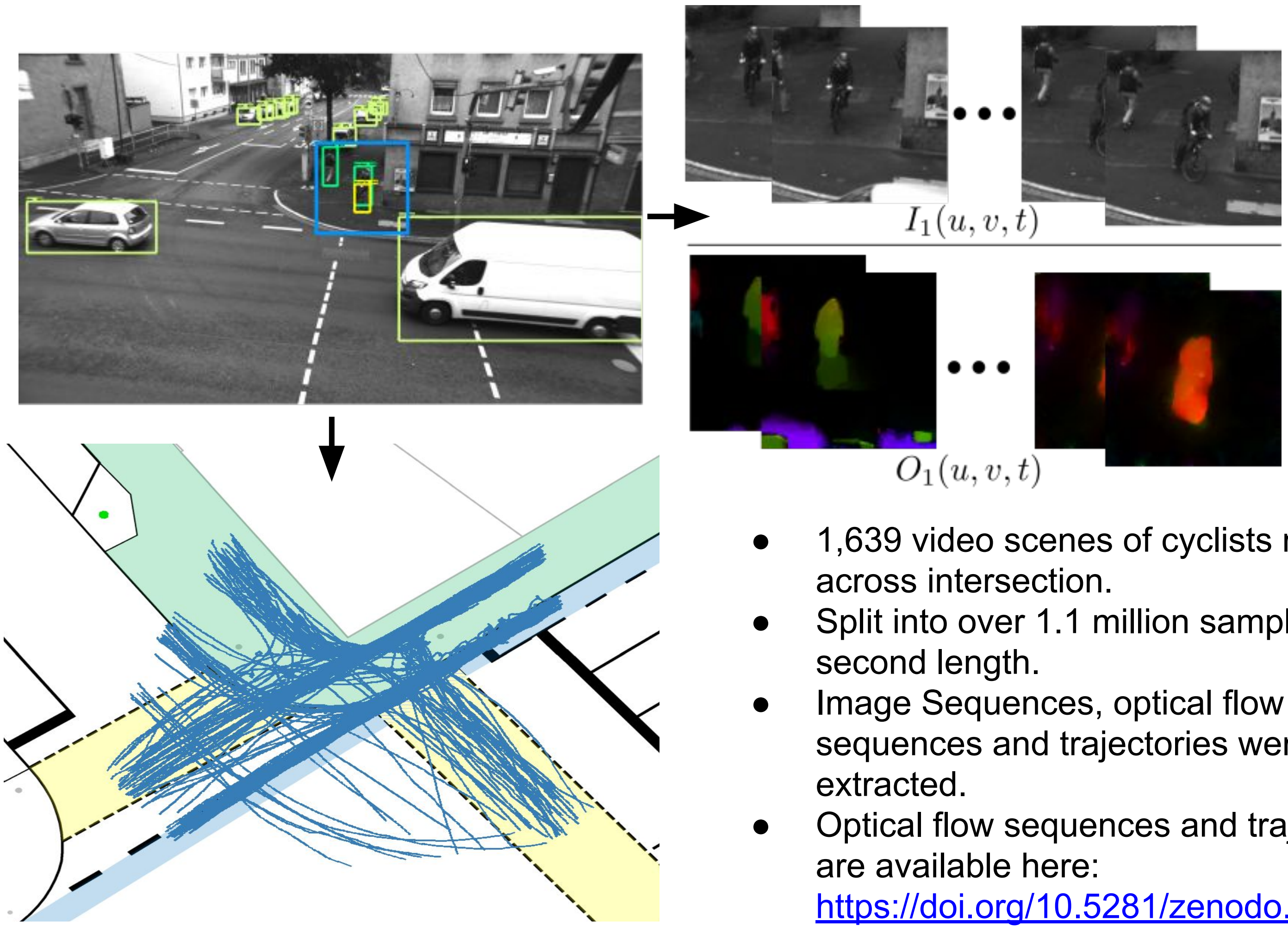


- Identify cyclist motion states at all times using video data.
- Detect transitions between motion states as early as possible.
- Create reliable state estimates.

Research Intersection

TH Aschaffenburg
university of applied sciences

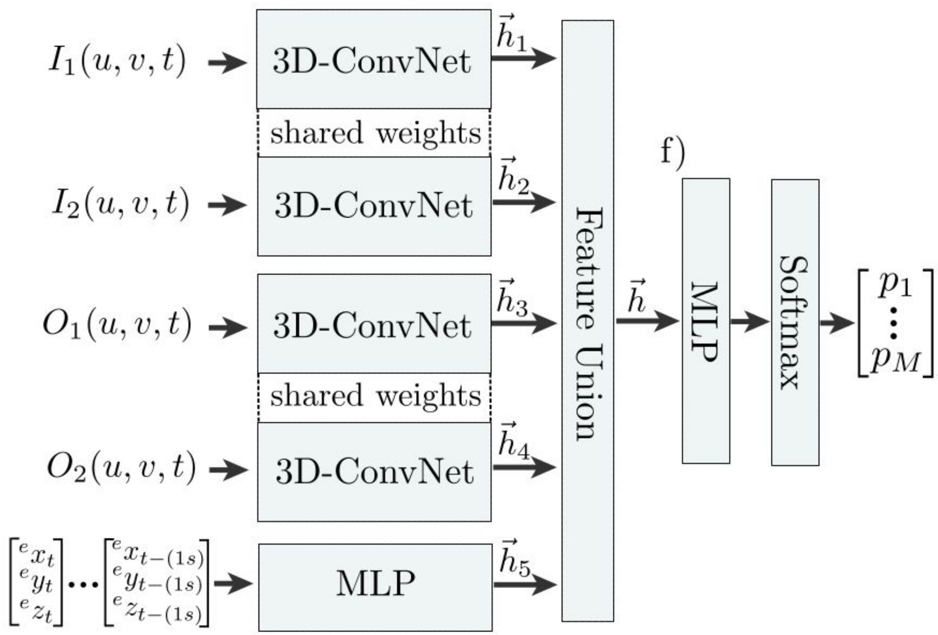
Dataset

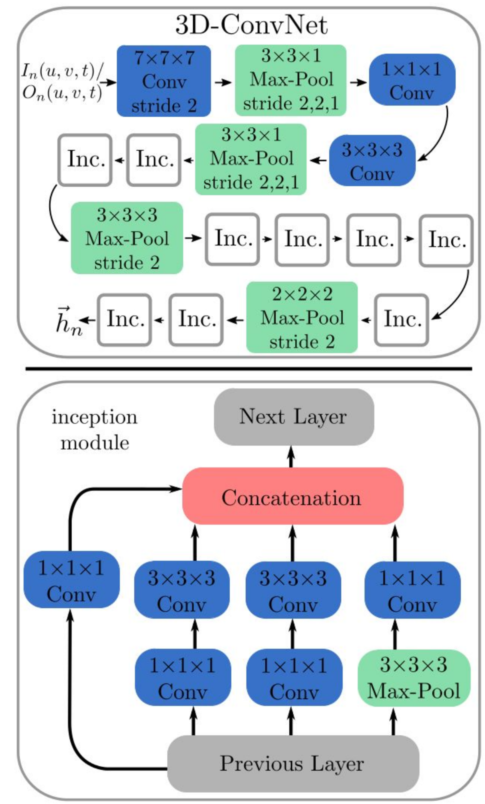


- 1,639 video scenes of cyclists moving across intersection.
- Split into over 1.1 million samples of 1 second length.
- Image Sequences, optical flow sequences and trajectories were extracted.
- Optical flow sequences and trajectories are available here: https://doi.org/10.5281/zenodo.3734038

TH Aschaffenburg
university of applied sciences

Method





- Multi-stream architecture using image sequences, optical flow sequences, and trajectories.
- ConvNets use Deepmind's I3D architecture [1].
- Motion state is classified in every time step.

[1] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 4724–4733.

Results

baseline model									
	wait/motion		*turn/straight*		*left/right*		*start/stop/move*		
$F_{1,seg}$	0.813		0.550		0.909		0.400		
	wait	*motion*	*turn*	*straight*	*left*	*right*	*start*	*stop*	*move*
$F_{1,seg}$	0.604	0.878	0.491	0.569	0.956	0.863	0.311	0.557	0.390
$\bar{t}_d$	0.081 s	0.062 s	0.180 s	0.063 s	0.012 s	0.029 s	0.033 s	0.265 s	0.157 s

MS-Net									
	wait/motion		*turn/straight*		*left/right*		*start/stop/move*		
$F_{1,seg}$	0.825		0.697		0.932		0.567		
	wait	*motion*	*turn*	*straight*	*left*	*right*	*start*	*stop*	*move*
$F_{1,seg}$	0.635	0.884	0.431	0.761	0.908	0.954	0.312	0.497	0.656
$\bar{t}_d$	0.060 s	0.032 s	0.217 s	0.036 s	0.015 s	0.013 s	0.011 s	0.509 s	0.071 s

- Compared to a baseline model (trajectory only), the method using motion sequences leads to more accurate predictions and faster detection of transitions between motion states.
- The use of optical flow sequences alone leads to similar results compared to a model using motion sequences, optical flow sequences, trajectory inputs.
- The inference time of the model was measured at 41 ms.

Thank You

Thank you for watching!

I hope to talk to you at the poster presentation!