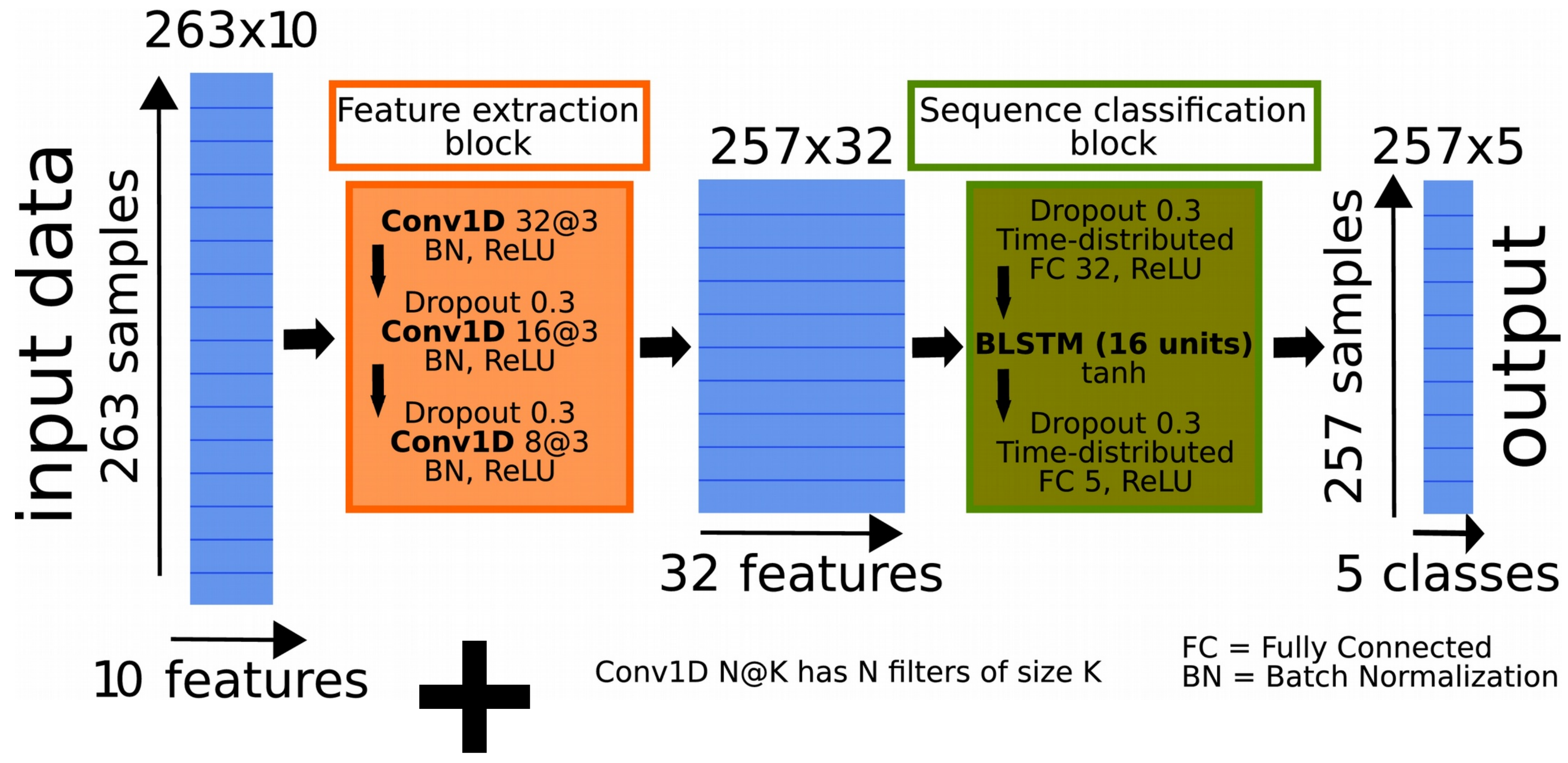
Deep Learning vs. Manual Annotation of Eye Movements

Mikhail Startsev[1], Ioannis Agtzidis[2], Michael Dorr[3]
Technical University Munich
[1] mikhail.startsev@tum.de, [2] ioannis.agtzidis@tum.de, [3] michael.dorr@tum.de

Deep Learning models have revolutionized many research fields already. However, the raw eye movement data is still typically processed into discrete events via threshold-based algorithms or manual labelling. In this work, we describe a compact 1D CNN model, which we combined with BLSTM to achieve end-to-end sequence-to-sequence learning. We compare the performance of our approach to various literature models and manual raters. Our deep method demonstrates superior performance, which brings us closer to human-level labelling quality.

Automatic Labelling



Conv1D N@K has N filters of size K

FC = Fully Connected
BN = Batch Normalization

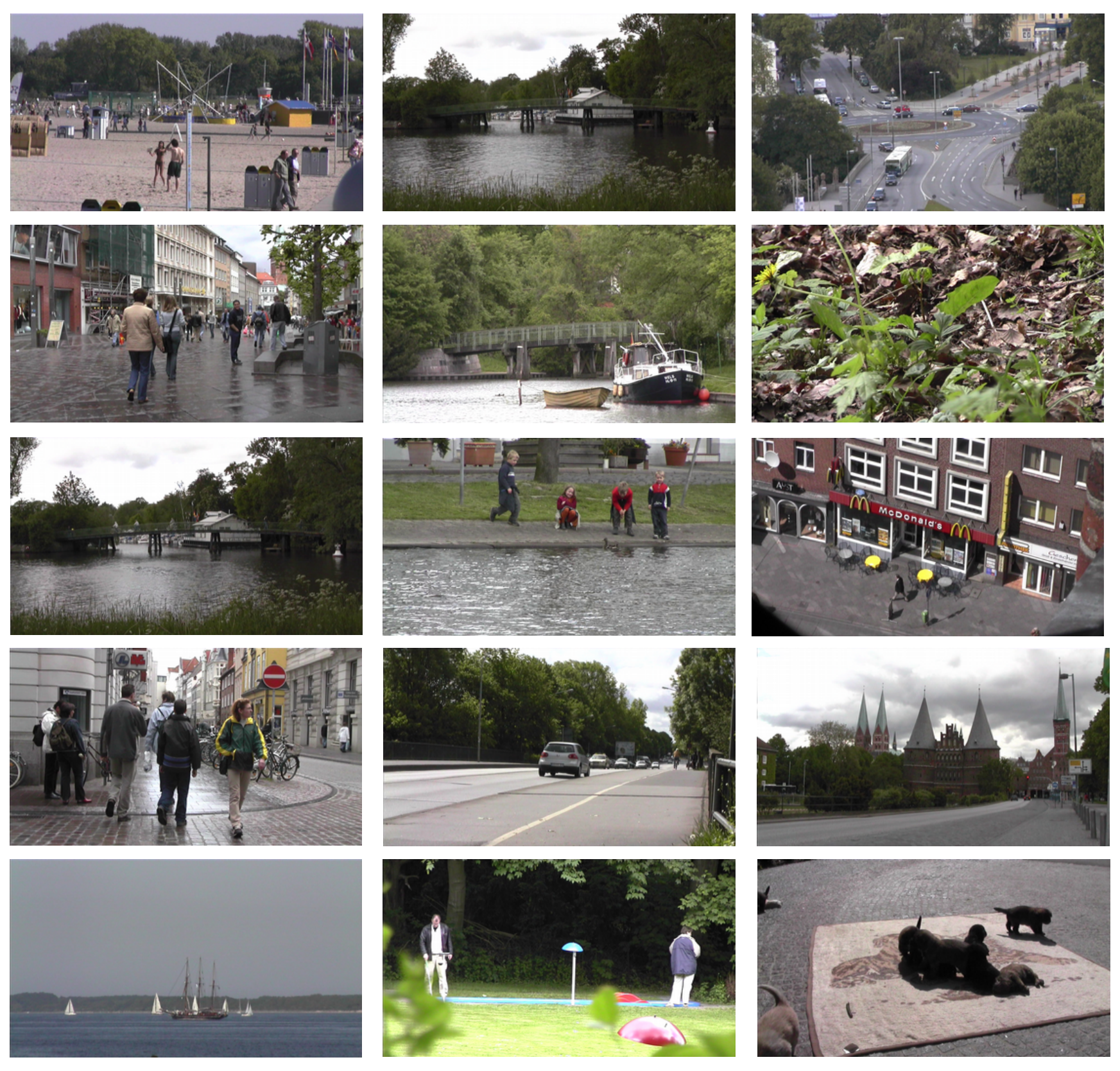
+ → < 1.5s per
1m of recording
(at 250Hz)

Compared to the state of the art on fully manually annotated GazeCom data set:
(CNN-BLSTM model evaluated via cross-validataion; parameters of I-VMP were optimized for best results on the entire data set – prone to overfitting)

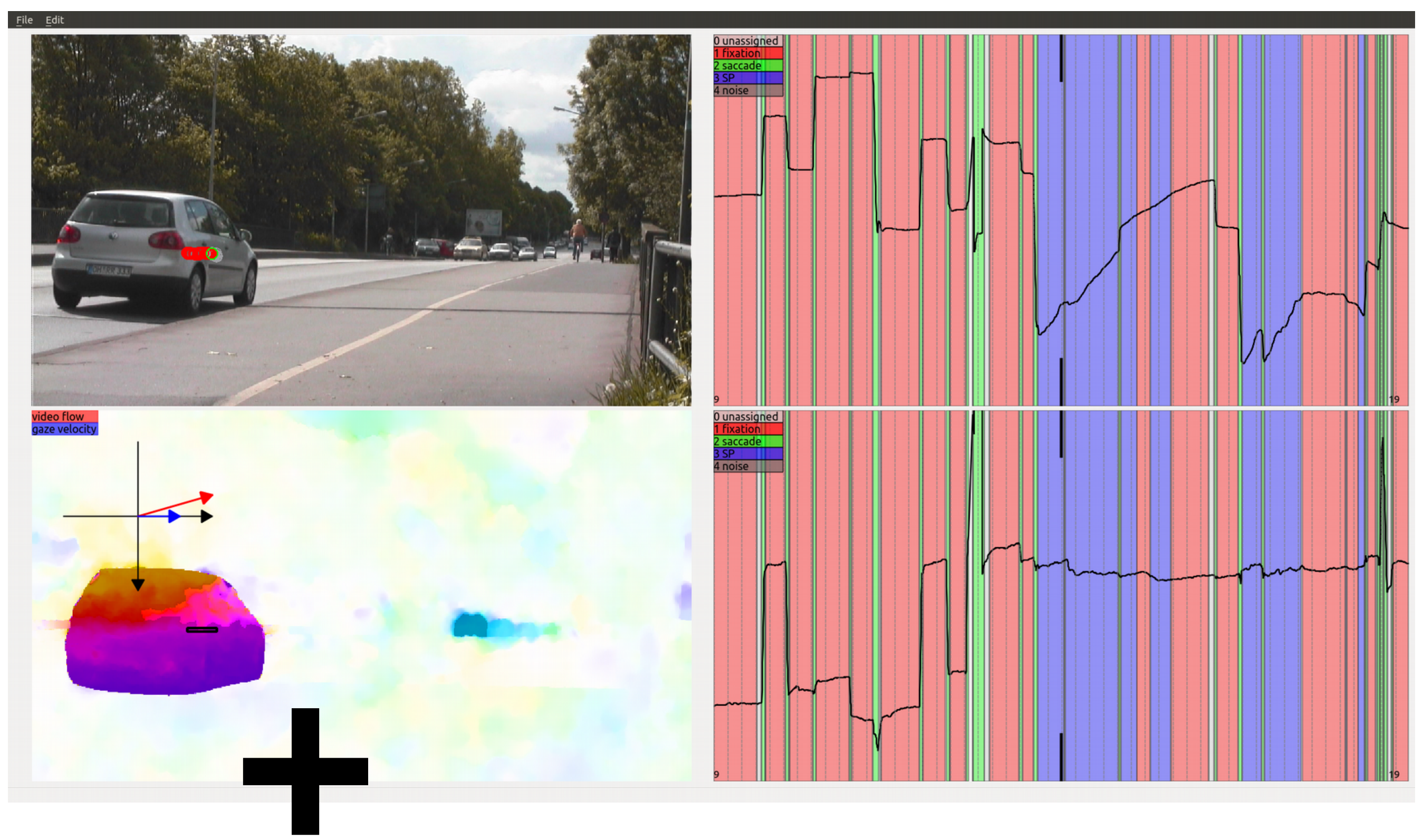
Model	***Fixation F1***	***Saccade F1***	***Pursuit F1***
CNN-BLSTM	**0.939**	**0.893**	**0.703**
[Agtzidis et al. 2016]	0.886	0.864	0.646
[Larsson et al. 2015]	0.912	0.861	0.459
I-VMP (optimistic)	0.907	0.725	0.570
[Dorr et al. 2010]	0.919	0.829	0.381
[Berg et al. 2009]	0.883	0.697	0.422

Our model shows consistently higher performance than any other literature approach we tested, including the multi-observer approach of [Agtzidis et al. 2016], which aggregates information across multiple eye tracking recordings for each video clip in order to produce smooth pursuit labels.

GazeCom data set:
- 18 clips
- 20s each
- over 4.5 viewing hours in total
- 47 observers per video on average
- full manual annotation of eye movements



Manual Labelling



+ → ca. 30 min per
1m of recording

We asked several additional experts to label a 5-second excerpt from the beginning of a GazeCom recording, where our model shows its median performance. For this data subset, we compare both the automatic models and the manual raters to our original ground truth:

Model	***Fixation F1***	***Saccade F1***	***Pursuit F1***
CNN-BLSTM	0.917	**0.863**	0.853
Expert 04	**0.934**	0.768	**0.905**
[Agtzidis et al. 2016]	0.886	0.809	0.871
Expert 03	0.927	0.720	0.895
Expert 02	0.877	0.776	0.765
Expert 01	0.805	0.831	0.748
I-VMP (optimistic)	0.814	0.722	0.696
[Berg et al. 2009]	0.703	0.570	0.437
[Dorr et al. 2010]	0.757	0.794	0.089
[Larsson et al. 2015]	0.744	0.790	0.0

Our model especially excels at learning the saccade labels, surpassing all the manual raters. Overall, the performance of our model is still very high.

The code, models, and data used for this work will be made available at http://michaeldorr.de/smoothpursuit