Optimizing Clustering-based Smooth Pursuit Detection

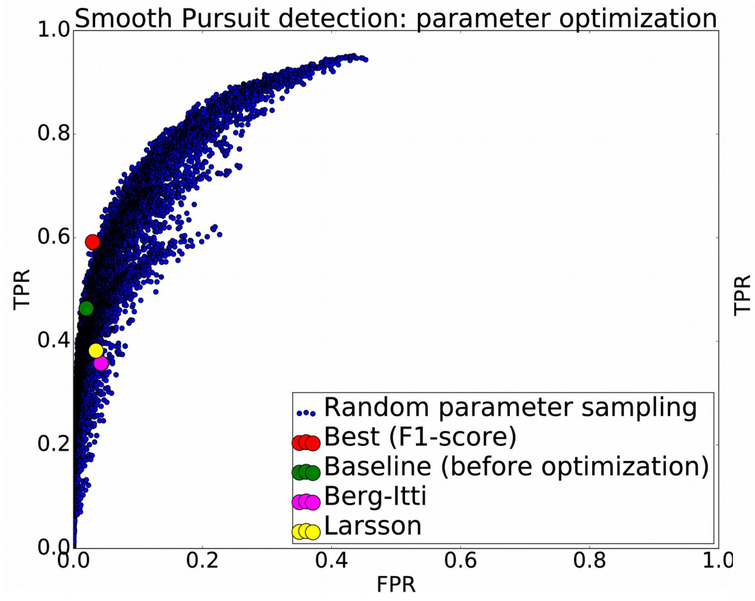
Mikhail Startsev[1], Albert Tae-Young Lee[2], Michael Dorr[3]
Technical University Munich
[1] mikhail.startsev@tum.de, [2] albert.lee@tum.de, [3] michael.dorr@tum.de

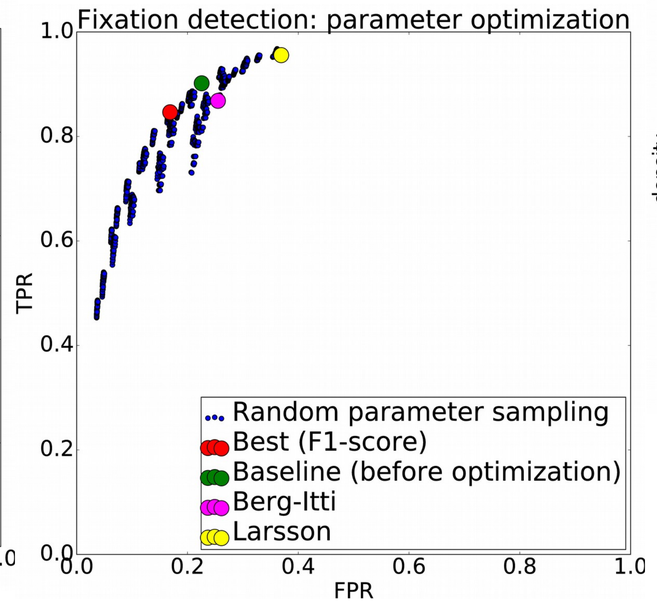
We investigate gaze behaviour with dynamic natural scenes, with an emphasis on smooth pursuit (SP). We previously developed an algorithm that leverages inter-observer similarity by clustering dynamic point-of-regard samples of multiple people. Without any parameter tuning, this approach outperformed the state-of-the-art algorithms, although it often produced fragmented SP episodes.

In this work we performed a cross-validated grid search for optimal parameters in log-space and, to tackle fragmentation, trained a Hidden Markov Model to implicitly smoothen detected SP episodes. Our previous algorithm achieved 57.1% SP-F1, 90.8% fixation-F1 (where F1-score is the harmonic average of precision and recall). After applying the modifications described in this work, we are able to reach 64.9% SP-F1 (at 88.7% fixation-F1), with the added benefit of a more natural SP episode duration distribution (Kullback–Leibler divergence down from 1.385 to 0.358).

I. Parameter Optimization

We perform a randomised grid search in log-space for all the fixation- and SP-detection parameters:





In total, over 11,000 parameter sets were sampled. We look for a “robust” well-performing set of parameters, i.e. it should be ranked highly on both subsets of the videos (in top-15 already 2 matching sets).

II. HMM Smoothing

The manually annotated data has mean SP duration of 0.41s. Our non-optimized SP detector produces SP episodes with mean duration of only 0.17s, which becomes 0.25s after optimization. To make the duration distributions match better, we implicitly smoothen our fragmented episodes by training a Hidden Markov Model, where observations are taken directly from the outputs our SP detector, while internal states match the labels of the manually annotated ground truth.

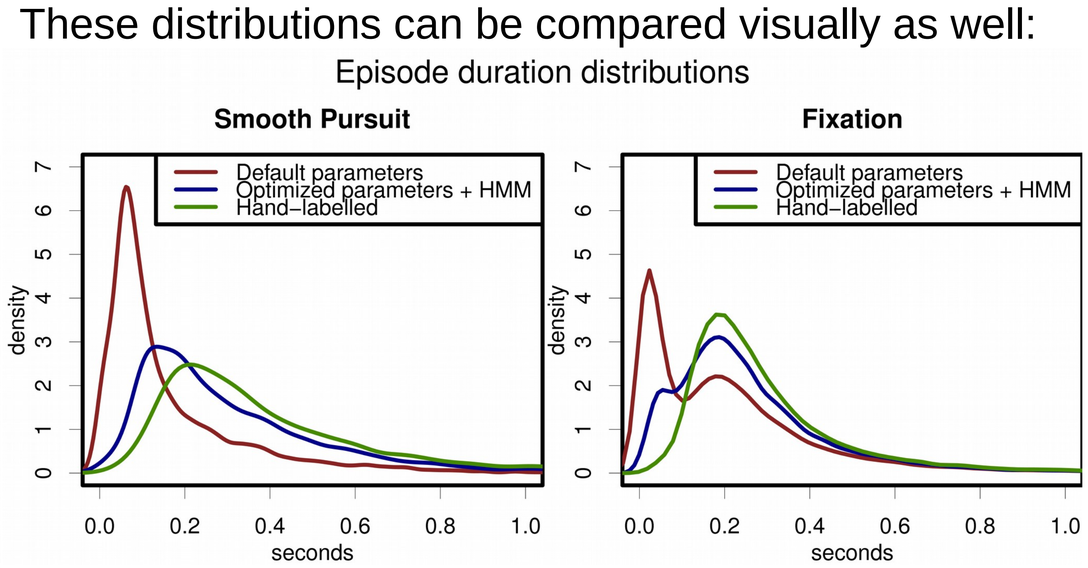
III. Evaluation & Results

all evaluation is performed via 2-fold cross-validation

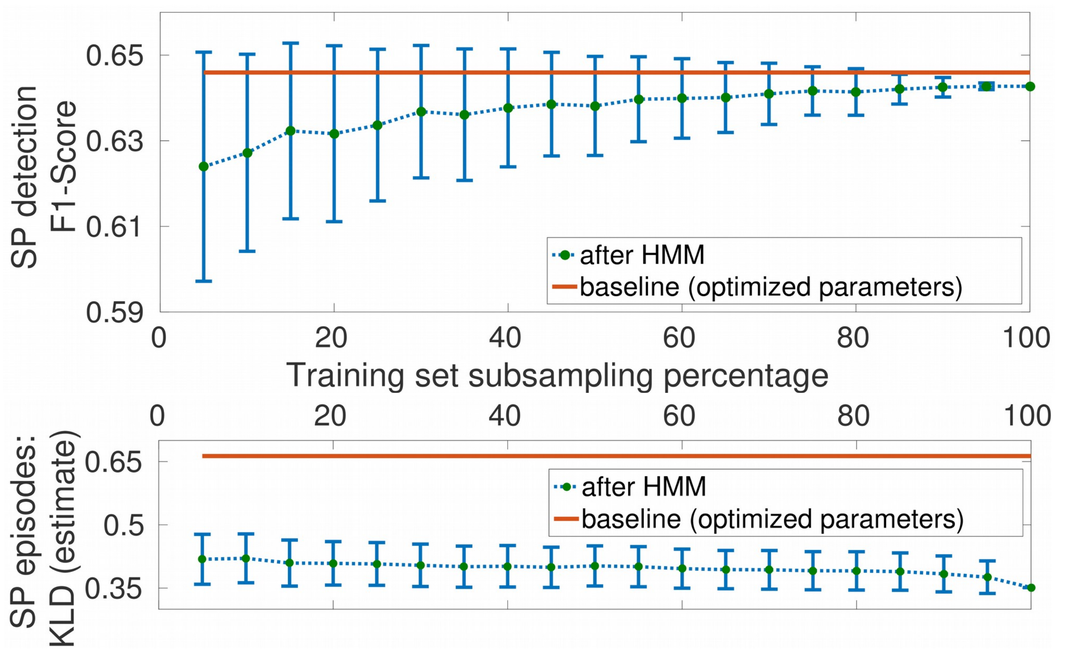
We measure the similarity between the distributions of durations of those SP and fixation episodes that were detected algorithmically and those in the hand-labelled data, whilst also comparing the detection power:

	KLD *lower is better*	**Histogram intersection** *higher is better*	**F1-score** *higher is better*
Default parameters	SP: 1.38 Fix.: 0.44	SP: 0.44 Fix.: 0.66	SP: 0.571 **Fix.: 0.908**
Default parameters + HMM	SP: 0.40 **Fix.: 0.16**	SP: 0.74 **Fix.: 0.86**	SP: 0.581 Fix.: 0.905
Optimized parameters	SP: 0.62 Fix.: 0.34	SP: 0.68 Fix.: 0.75	SP: 0.646 Fix.: 0.886
Optimized parameters + HMM	**SP: 0.36** Fix.: 0.18	**SP: 0.82** **Fix.: 0.86**	**SP: 0.649** Fix.: 0.887

These distributions can be compared visually as well:



We also observe that the size of the HMM training set does not have a huge impact on performance:



The code (both the original SP detector, and the HMM label smoothing tool) and the data used for this work can be found at http://michaeldorr.de/smoothpursuit

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