Supplementary Material: An Intent-Based Automated Traffic Light for Pedestrians

Christian Ertler Horst Possegger Michael Opitz Horst Bischof
Institute of Computer Graphics and Vision, Graz University of Technology, Austria
{christian.ertler,possegger,michael.opitz,bischof}@icg.tugraz.at

In the following, we provide additional evaluations to extend our ablation study in Section 1 and present further insights into our long-term study in Section 2.

1. Additional motion model evaluations

The accuracy (Acc) evaluation in the main paper (Section 4.2.) is based on the percentage of frames with correct exit predictions of each trajectory. By thresholding this measure, we can compute the number of *True Positives (TP)*, *True Negatives (TN)*, *False Positives (FP)* and *False Negatives (FN)* and finally compute the accuracy as

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

While this is a standard measure for classification problems, it neglects an important factor for our application: the duration between the first occurrence of a pedestrian until a reliable intent prediction can be made. Our goal is to decide about the intent as soon as possible to make for fast traffic light scheduling and thus, improve traffic flow.

To this end, we introduce a quality measure for the task of intent prediction, which we call LAST POINT OF WRONG DECISION (LPWD). Intuitively, this measure denotes the last time step the intent prediction for a trajectory was wrong. More formally, we define

$$\begin{aligned} \text{LPWD}_k = \max_{j \in [1, |t_k|]} \frac{j}{|t_k|}, \\ \text{s.t.} \quad p_{k,\tilde{r}_k,m}^{(\alpha_k + j)} \leq 1/R, \end{aligned} \tag{2}$$

where $|t_k|$ is the length of trajectory t_k, α_k is the birth time of t_k, $\tilde{r}_k \in [1, R]$ is the correct exit region for t_k and $p_{k,\tilde{r},m}^{(i)}$ with $m \in \{\text{EX}, \text{GM}, \text{COMB}\}$ are the model component predictions of our system. Further, we take the mean over all trajectories

$$\text{mLPWD} = \frac{1}{K} \sum_{k=1}^{K} \text{LPWD}_k, \tag{3}$$

Method	mLPWD	mLSWD
EX MODEL	0.851	14.050
GM MODEL	0.209	2.060
COMBINED MODEL	**0.185**	**1.930**

Table 1. Predicting the intent (*i.e.* correct exit region) with different motion model components. **mLPWD** and **mLSWD** are the mean LAST POINT OF WRONG DECISION and LAST SECOND OF WRONG DECISION, respectively; lower values indicate better performance. The mean lifetime of a trajectory in the dataset amounts to 15.260 s.

where K is the total number of trajectories in the dataset. This measure allows to compare two methods in terms of the time it took to arrive at a reliable intent prediction; a lower value is better. While LPWD_k is relative to the trajectory length normalized between 0 and 1, we also report the LAST SECOND OF WRONG DECISION (LSWD_k), which encodes the same information in terms of seconds passed since the birth of trajectory k.

Table 1 shows additional ablation results for the motion models defined in the main paper. Again, the GM MODEL performs better than the EX MODEL and the COMBINED MODEL performs best. In Figure 2 we show qualitative examples of trajectories and the corresponding predictions illustrating strengths and weaknesses of each motion model.

2. Long-term study

In this section we provide more details on the comparison of our system to traditional, manual push-button solutions. We first elaborate on how we match the button triggers to our automated triggers; then, we provide a manual evaluation of our system on a subset of our data; finally, we show an analysis of dominant walking directions in our crosswalk scenes.

2.1. Vision-based versus push-button

A crucial part of comparing push-button systems to corresponding automated triggers is the matching. We record the timestamps of every manual push-button trigger and ev-

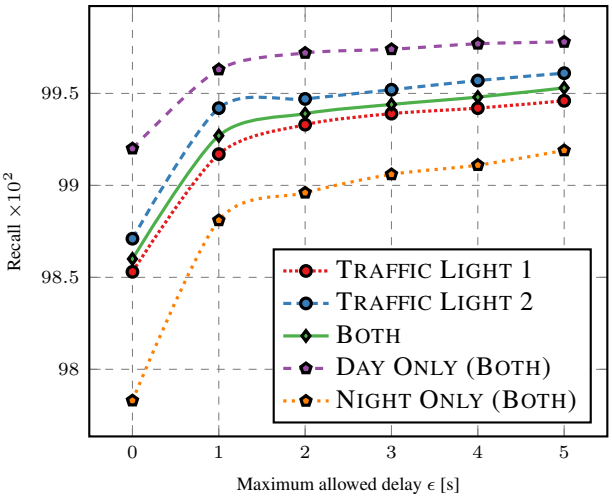


Figure 1. Recall over maximally allowed delay ϵ between push-button triggers and pedestrians reported by our system. TRAFFIC LIGHT 1 and TRAFFIC LIGHT 2 denote the two sides of the crosswalk, DAY ONLY includes triggers from 8am–3pm only and NIGHT ONLY the rest of the day.

ery trigger of our system. For each push-button trigger, we search for reported crossing requests of our system within a time span of at most ϵ seconds after the push-button trigger and 5 s before it. In this way, we can match both in-time and delayed predictions of our system, while we show in the main paper (Section 4.3; Figure 3) that the majority of predictions is not delayed – in fact, we show that we can make predictions severals seconds before push-button triggers.

In Figure 1 we show the recall of matched push-button triggers over different delays ϵ, where $\epsilon = 0$ means that we do not allow any delay of prediction. Even in this strict case we achieve high recall values. We see that the two independent systems at each end of the crosswalk perform comparably. The performance gap between daytime and nighttime was expected due to the harder illumination conditions for the object detector as our prototype did not use active illumination at night. However, our system is still able to reliably predict crossing requests even in this situation.

2.2. Manual evaluation

In addition to the automated comparison to the push-button, we also conducted a manual evaluation. It would have been infeasible to manually check all recorded pedestrians and push-button triggers both in terms of time and disk space, since we need to inspect series of images in order to judge the system's decision about a situation. Instead, we periodically (every 30 min) saved such image series together with the system state and predictions. Each of these situations was manually classified by a human annotator into:

	#{samples}	Recall ↑	Precision ↑
TRAFFIC LIGHT 1	1,578	0.992	0.972
TRAFFIC LIGHT 2	1,569	0.995	0.962
BOTH	3,147	0.994	0.967

Table 2. Manual evaluation of periodically sampled situations. *Traffic light 1* and *Traffic light 2* denote the two sides of the crosswalk.

TP: At least one pedestrian wants to cross the road and the system reported a crossing request correctly.

TN: No pedestrian wants to cross the road and the system did not report a crossing request correctly.

FP: No pedestrian wants to cross the road and the system reported a crossing request by mistake.

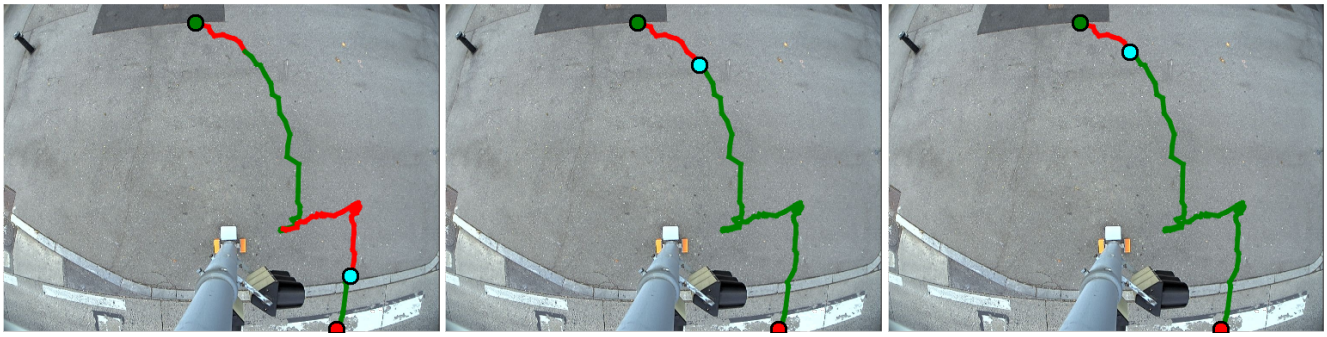
FN: At least one pedestrian wants to cross the road and the system did not report a crossing request by mistake.

We report the results together with the absolute numbers in Table 2. The recall is consistent with the findings in Section 2.1; additionally we find high precision of > 0.96 for both sides of the crosswalk. Most of the **FN** samples are due to bad illumination conditions at night; **FPs** happen primarily during snowy or rainy weather, where the detector yields false positive detections and – to a lesser extent – due to wrong intent predictions.

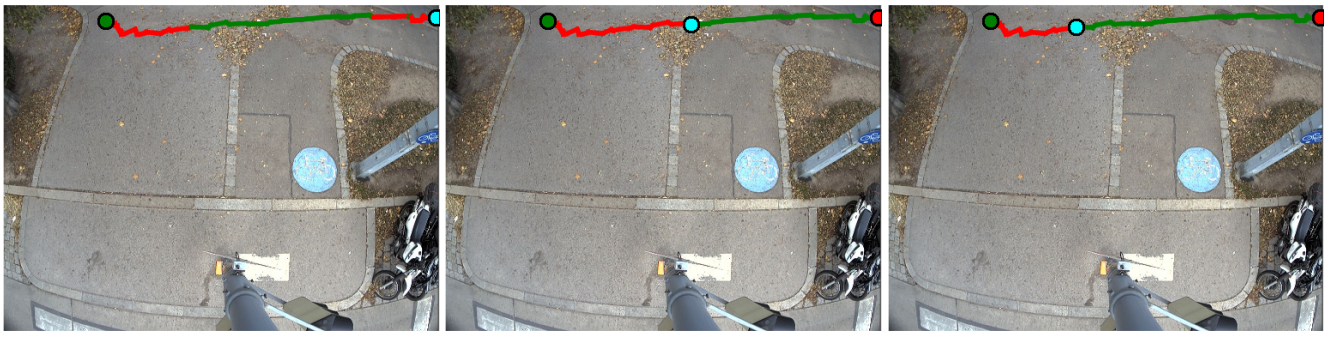
2.3. Analysis of walking directions

Figure 3 shows camera inputs with overlays of the dominant walking directions. We select these directions by first clustering the trajectories in our dataset by the location of their initialization and termination. We then visualize each cluster by its mean trajectory combined with the upper and lower quartiles to indicate the motion variance within the cluster. To avoid unnecessary visual clutter, we only show the clusters covering the majority of the trajectories.

This visualization of pedestrian movements also demonstrates the difficulty of correctly predicting a pedestrian's intent. For example, consider TRAFFIC LIGHT 1 in Figure 3 (left): Will a person who enters the field-of-view at the top-right corner proceed to pick up her car (at the parking lane) or move towards the crosswalk? In such scenarios even a human operator would have to delay a prediction and continue to observe the pedestrian's trajectory until enough evidence for a particular exit region could be gathered.



(a) A pedestrian walks towards the push-button and all models yield correct predictions; she waits for the green phase and walks around, resulting in instability of the EX MODEL; it becomes stable again when the pedestrian heads towards the road. The COMBINED MODEL chooses the best prediction in all phases.



(b) A pedestrian passes by and does not want to cross the road. The EX MODEL is faster than the GM MODEL at predicting the correct intent in the beginning but fails at the end due to noisy bounding box detections at the image border. The COMBINED MODEL fuses the complementary information and yields the best decision at all time steps.

Figure 2. Visualizing predictions for exemplary trajectories (best viewed in color). The start and end points of the trajectory are shown as ● and ●, respectively; the LAST POINT OF WRONG DECISION (LPWD) is visualized with ●. Green trajectory segments indicate that the intent prediction was correct at the corresponding time steps; red segments denote wrong predictions. **Left:** EX MODEL. **Middle:** GM MODEL. **Right:** COMBINED MODEL.

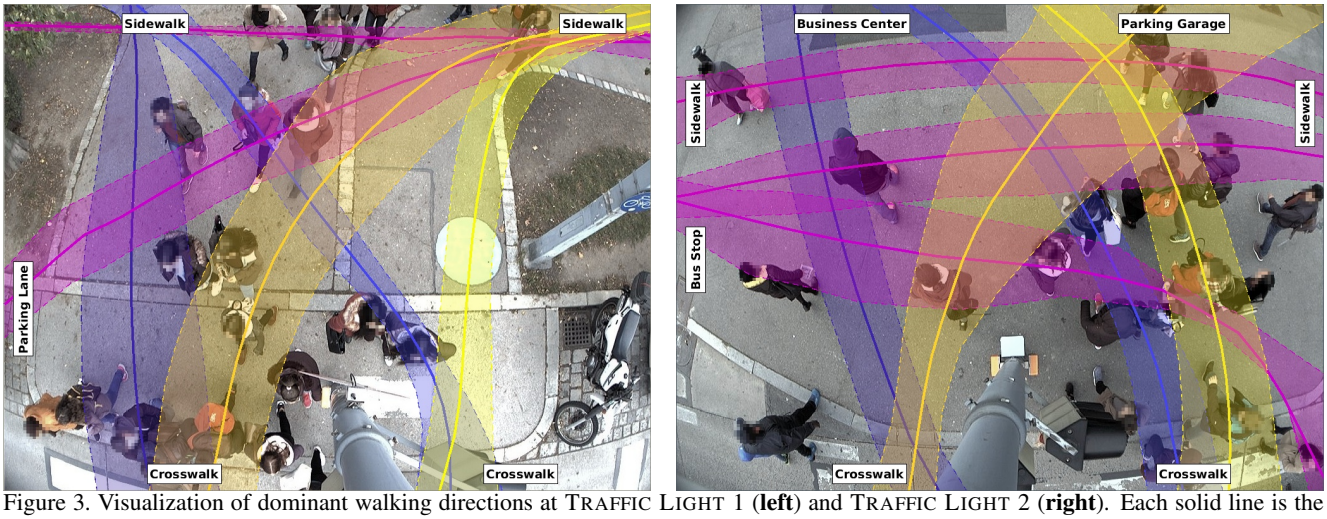


Figure 3. Visualization of dominant walking directions at TRAFFIC LIGHT 1 (**left**) and TRAFFIC LIGHT 2 (**right**). Each solid line is the mean trajectory of a cluster; the dashed lines show the upper and lower quartiles. Nearby points of interest are indicated by the text boxes. Best viewed in color.