

Promotion of Proactive Behavioral Change to Prevent Human Error

- Part 5 Development of a Fundamental Technology for Integrating Attention Targets Extracted via LLM-based Traffic Context Understanding with Driver Gaze Targets-

Yamasaki Soma ¹⁾ Yamasaki Akito ¹⁾ Yoshitake Hiroshi ²⁾ Shino Motoki ²⁾

1) Meijo University

501-1 Shiogamaguchi, Tenpaku, Nagoya, Aichi, 468-8502, Japan

2) Institute of Science Tokyo

2-12-1 Ookayama, Meguro, Tokyo, 152-8550, Japan

KEY WORDS: Human engineering, Human error, Driving act/driver behavior, Driver attention, Recognition/judgment, Driving support/driver support, Large Language Model(C2)

Many traffic accidents stem from insufficient confirmation behavior by drivers; for example, failing to check for pedestrians or vehicles hidden in a blind spot at a poor-visibility intersection leads to a collision. Reducing such accidents requires proactive behavioral change, in which drivers themselves become aware of their inadequate confirmation and improve how they drive. This is especially important for elderly drivers, whose accidents are increasing and who tend not to recognize their own decline, so that support for awareness based on objective evidence is needed. To promote such change, it is necessary to present what the driver failed to look at, together with concrete grounds. This requires describing not only what the driver did, but also what should have been looked at in that scene—the requirement imposed by the traffic environment—which has not yet been established.

Conventional environment recognition using object detection or semantic segmentation relies on predefined categories, and although it can describe what is where, it cannot handle context-dependent requirements such as "slow down and check both sides at a poor-visibility intersection." A large language model (LLM), by contrast, has broad knowledge linking images and language and can describe the meaning of a situation that is not defined in advance as a category. This study therefore uses an LLM to understand the traffic context from forward-view images and to extract the requirements of the traffic environment as attention targets, converting them into structured data with importance levels and image regions. Integrating these targets with the driver's gaze, previously visualized by our method that estimates head pose and gaze from driver images, provides a basis for quantitatively evaluating confirmation behavior.

In the proposed method, a sequence of frames from the forward-view video is input to the LLM and treated as one continuous scene, so that changes such as an approaching pedestrian or a changing signal are considered(Fig.1). Priority is given in the order of regulatory elements, dynamic elements, road geometry, and environmental factors, and after a stepwise reasoning process each target is output in a structured form with its reason and importance. The targets are divided into *fixation_targets*—concrete objects to be visually confirmed, such as signals and pedestrians—and *caution_areas*—structural or situational elements that raise potential hazards, such as blind spots created by parked vehicles. The *fixation_targets* are then localized as bounding boxes by open-vocabulary object detection(Fig.1); since the LLM identifies the hazardous individual from context, the intended object is selected using its positional description when several objects match one query. For *caution_areas*, the object creating the blind spot is detected and an attention region is placed on the side where a pedestrian may emerge.

The extracted targets are integrated with the driver's gaze to evaluate confirmation behavior. For a *fixation_target*, it is judged as confirmed if the gaze enters its bounding box at least once during the scene, missed if it never does, and excluded if no box is obtained. The main index is the confirmation rate—the summed weight of confirmed targets divided by that of all targets—where each weight is the product of a category weight and an in-category rank weight. A list of critical misses among high-importance targets is also provided. For caution areas, the overlap between the required-attention heatmap and the driver's gaze-distribution heatmap is evaluated as a continuous score, kept separate from the confirmation rate. Future work will verify whether the extracted attention targets and their importance agree with what should actually be attended to, by comparison with the judgment of experts such as driving instructors, and then apply the method to real driving data to examine its effect on behavioral change.



"name": "red triangular stop sign small red triangular stop sign"
"reason": "It requires the driver to come to a complete stop before entering the intersection."
"rank":1
"name": "blue circular direction sign"
"reason": "It gives mandatory direction information that affects how the driver may proceed through the intersection."
"rank":2
"name": "blue pedestrian crossing sign"
"reason": "It highlights the presence of the crosswalk and the need to watch for pedestrians."
"rank":3
"name": "zebra crosswalk"
"reason": "Pedestrians may be crossing here, so the driver must slow and be prepared to yield before the stop line."
"rank":4
"name": "convex traffic mirrors"
"reason": "They indicate limited sight lines and help the driver check for cross traffic or pedestrians approaching from hidden areas."
"rank":5

Fig. 1 Example of attention targets detected on a forward-view image. Each box shows an object name and its detection confidence.