Available online at www.sciencedirect.com
ScienceDirect
Artificial Intelligence 172 (2008) 351–391
Artificial Intelligence
www.elsevier.com/locate/artint

# Representation of occurrences for road vehicle traffic

R. Gerber, H.-H. Nagel *

*Institut für Algorithmen und Kognitive Systeme, Universität Karlsruhe (TH), 76128 Karlsruhe, Germany*

Received 28 June 2004; received in revised form 17 July 2007; accepted 19 July 2007

Available online 27 July 2007

**Abstract**

Our 3D-model-based Computer Vision subsystem extracts vehicle trajectories from monocular digitized videos recording road vehicles in inner-city traffic. Steps are documented which import these quantitative geometrical results into a conceptual representation based on a Fuzzy Metric-Temporal Horn Logic (FMTHL, see [K.H. Schäfer, Unscharfe zeitlogische Modellierung von Situationen und Handlungen in Bildfolgenauswertung und Robotik, Dissertation, 1996]). The *facts* created by this *import step* can be understood as verb phrases which describe elementary actions of vehicles in image sequences of road traffic scenes. The current contribution suggests a *complete conceptual* representation of *elementary vehicle actions* and reports results obtained by an implementation of this approach from real-world traffic videos.


*Keywords:* Computer vision; Knowledge representation; Temporal reasoning; Reasoning about actions and change; Fuzzy metric-temporal logic

## 1. Introduction

An adult is expected to be able to write down—not necessarily with style and precision—what he sees. Conceding similar, but appropriately adapted reservations, what is required to have a computer perform an analogous task? Obviously, an analogue to human seeing could be Computer Vision. The notion of an automatic report generator, too, is no longer considered as science fiction. It most likely turns into a challenge, however, to imagine the detailed communication between a computer vision (sub)system and an algorithmic report generator. What looks like the mere definition of an interface will turn out to require the design of a system-internal *logic-based* conceptual representation of a text in combination with the design of an entire set of processes operating on this representation.

Investigations to be discussed in the sequel address an important step towards the algorithmic transformation of video signals into a natural language text which describes the recorded scene, in particular its temporal development. The presentation will first sketch an overall system concept in order to provide a framework for the subsequent discussion which will then concentrate on the conversion of *geometric tracking results* into *elementary conceptual representations* of relevant aspects of the (short-term) development in the recorded scene. A preliminary version of this approach has been partially outlined in [10].

* Corresponding author.
*E-mail address:* nagel@iaks.uni-karlsruhe.de (H.-H. Nagel).

0004-3702/$ – see front matter 
doi:10.1016/j.artint.2007.07.001

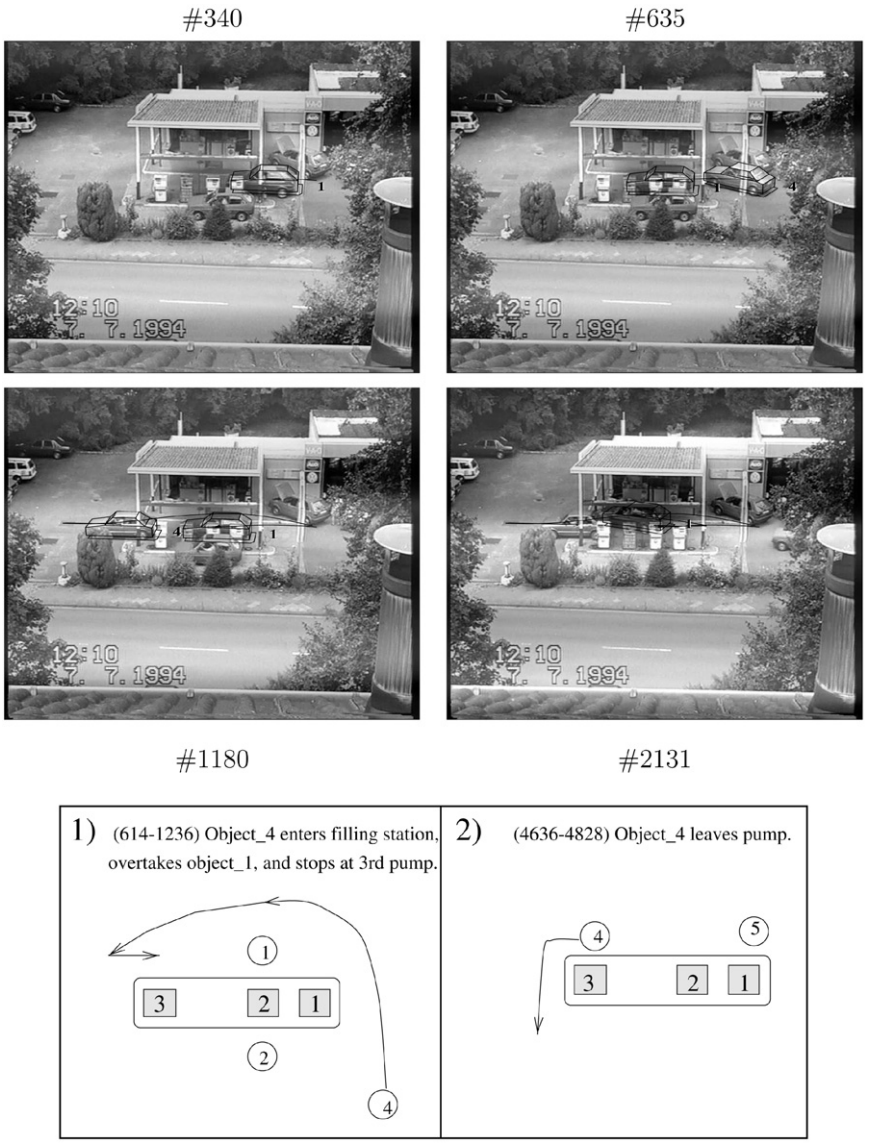


Fig. 1. In the upper left panel, the image plane projection of a polyhedral model for a fastback has been overlaid to frame number 340 from an image sequence recorded at a gas station. In addition, one can see the trajectory segment obtained by automatic model-based tracking of this vehicle which will be referred to as *object_1*. Frames 635, 1180, and 2131 show snapshots of various maneuvers of another vehicle (*object_4*), with analogous overlays of a projected polyhedral model and the trajectory for this vehicle (see Section 1 for more explanations). The sketch in the bottom row illustrates the maneuvers of object_4 in this sequence.

Fig. 1 illustrates a coherent source of examples for different stages of such a transformation. The first frame[1] #340 in the upper left panel shows a snapshot where a fastback has already stopped at the second petrol pump on the filling lane (see Fig. 2) closer to the observer (subsequently referred to as the 'lower filling lane'). A second fastback (subsequently referred to as 'object_1') had just entered the gas station and selected the filling lane on the other side

[1] Based on special derivative operators which suitably interpolate between digitizations in even and odd scanlines of interlaced video (see, e.g., [32]), actually each half-frame—or field in video-coding terminology—is evaluated in its own right, resulting in a temporal sampling rate of 20 msec. This aspect reduces the approximation errors by the Extended Kalman-Filter used and thus improves the tracking quality. Beyond this fact, however, it does not influence the conversion of geometric results to natural language concepts. In order to simplify the presentation, we shall use the term frame henceforth without further qualifications.