On the Use of Motion Concepts for Top-Down Control in Traffic Scenes

Michael Mohnhaupt, Bernd Neumann
University of Hamburg, FB Informatik

1. Introduction

The use of models for top-down control is the major strategy to beat the inherent complexity of many visual processes. Fortunately, information to select appropriate models from long-term memory for top-down control is often available for a visual system. For example, it can be provided by previous bottom-up analysis, by the spatio-temporal context, by expectations about a scene, by goals or intentions of the system, and by other information sources, e.g., prior descriptions in natural language and the like. The main research focus in model-based vision has been on the use of static information, for example, the use of object models (see e.g. [Tsotsos 87] for an overview).

In this paper we concentrate on temporal aspects of model-based vision: the use of motion concepts for top-down control. Motion concepts can constrain visual processes in two ways. First, they can provide a spatial focus for analysis, because instances of motion concepts can typically only be found at certain locations in a scene (e.g., a 'turn-off' event can only take place at intersections). Second, motion concepts can focus the analysis on a specific spatio-temporal behavior. We investigate both aspects in the domain of street traffic scenes, where typical objects are cars, pedestrians, trucks, and so on, and typical motion concepts are 'turn-off'-events, 'overtake' events, 'cross'-events and the like. In our examples, top-down information is given by natural language utterances.

Two central and interrelated questions are: 1) At which level of representation should bottom up processes and top-down control interact? and 2) How should motion concepts be represented to support top-down guidance? In [Mohnhaupt + Neumann 90] we propose a hybrid representation of motion concepts. A propositional representation including a logic based style of reasoning is exploited for event recognition and long-term memory, and an analogical quantitative spatio-temporal buffer is used for motion visualization and prediction, for learning object motion and several aspects of spatio temporal reasoning. The buffer facilitates important tasks related to concrete visual scenes. It can be instantiated on demand from long-term memory.

Here, we focus on one aspect of the spatio-temporal buffer: the generation of predictions suited for top-down control of motion analysis. The central idea is to express motion concepts as typicality distributions in the buffer. The buffer is shared between bottom-up and top-down processes. We show that spatio-temporal constraints for motion analysis can be derived, and we sketch how models and bottom-up data can interact to allow for meaningful predictions. Both leads to a significant reduction of complexity: For traffic scenes top-down control through the use of motion concepts can reduce the amount of computation by several orders of magnitude.

2. Motion concepts implied by verbs of locomotion

In street traffic scenes we associate motion concepts with verbs of locomotion like 'drive', 'walk', 'turn-off', etc., as proposed by [Neumann 89] for bottom-up event recognition using propositional event models. These event models are inappropriate for top-down control mainly for two reasons: First, predictions in terms of predicates are unnecessarily imprecise, because typical and atypical instances cannot be distinguished. And second, propositional event models are difficult to adapt to constraints provided by bottom-up analysis, for example obstacles on the street; clearly this should lead to an adapted prediction.

This lead us to consider a spatio-temporal buffer representation for top-down control. It is shared by bottom-up and top-down processes and closely related to perceptual representations. The buffer is fourdimensional (x, y, direction of velocity, speed). It can be filled with a typicality field for motion in a certain subfield of the xy-plane in the scene, for example, a typicality field representing a turn-off model for a particular intersection. The typicality field results from accumulated and processed event instances (see [Mohnhaupt + Neumann 89], [Mohnhaupt + Neumann 90]). Stationary scene objects like the street shape can also be filled in, from model-based expectations as well as from

visual processes. The spatio-temporal buffer is an extension of the purely spatial buffer proposed by [Kosslyn 80]. One can think of it as an internal image-like representation with temporal behavior to simulate events of interest and to derive helpful information.

Given the typicality field of an event and a starting situation resulting from bottom-up analysis, a search space for the likely progression of the event instances can be computed by following all typicality values above a certain threshold. A subsequent motion analysis can focus on this search space which comprises the spatial and spatio-temporal constraints.

3. An example

Consider the task of analyzing a typical street traffic scene as depicted in Figure 1, a synthetic model of a real scene. The number of interesting objects and events, which could in principle be analyzed can be very big. To focus the analysis we assume top-down information given in terms of a natural language question like: *Did a car driving towards Dammtor turn off Schlüterstreet into Bieberstreet in front of the FB Informatik?* Top-down control is now performed in two steps. First, spatial constraints are exploited, and second additional motion constraints are derived.

Location information associated with a motion concept can be derived as follows. We assume the semantic content of the utterance to be represented in a case frame representation, including slots for agent, object, location goal, destination, directional, and the verb. The verb determines an event model (here the event model for 'turn-off'). From information about the applicability of turn-off events it is derived that they can only happen at intersections. The locative of the case frame allows to choose the intersection Schlüterstreet/Bieberstreet as a focus for analysis. In addition, the directional entry of the case frame further constrains the analysis, because a particular turn-off area can be inferred as shown in Figure 1 (dark area).

Our main focus is now on motion information associated with the motion concept to allow for further top-down control. Within the depicted dark area in Figure 1, a certain direction of motion and a certain speed can be expected in case of a turn-off event from Schlüterstreet into Bieberstreet driving towards Dammtor. Hence, the next step is to instantiate the spatio-temporal buffer with the scene geometry and the typicality distribution for turn-off. Given a starting point of the turn-off event in the image sequence, a spatio-temporal search area for subsequent motion analysis can be generated by considering continuations above a certain typicality. The prediction algorithm is local and provides location and velocity information. In our example all the successor cells with typicality values above a certain threshold lead to Figure 2.

In the example in Figure 2 the typicality distribution results from observing several turn-off examples, see Figure 3. xy-traces of observed examples are shown, objects are represented by their center of mass. Note that information about velocity is not visible in Figure 2 and Figure 3 but is part of the model. After recording examples, subsequent local processing leads to generalizations which cover the approximate area represented by the examples (see [Mohnhaupt + Neumann 90] for details of the learning steps and methods to instantiate typicality distributions from long-term memory and from models recorded in a different environment).

To support our considerations a real image sequence was recorded on this intersection. Figure 4 shows one frame of this sequence. The most interesting event for now is the white taxi turning off Schlüterstreet. Other moving objects include cars and pedestrians.

In order to be able to apply the constraining information shown in Figure 2 to the image sequence, the low-level motion representation is based on 3-dimensional Gabor cells [Adelson + Bergen 85]. The implementation is described in [Fleet 88]. The output of spatio-temporal Gabor cells is well suited for top-down control, as by using a Gabor filter bank an image sequence is decomposed into orientation, velocity, and scale specific information. Hence, top-down constraints can be brought to bear by selecting the appropriate subset of cells for an analysis (in the example, only those cells are chosen which are sensitive to motion towards the upper left according to the predictions computed within the buffer). Figure 5 shows the spatio-temporal energy of Gabor cells which are maximally sensitive to an orientation of 45 degrees with a speed of one pixel per frame. The main information about the taxi is within the predicted area, other motions as well as static information are removed.

4. Summary

We showed how to exploit motion concepts associated with verbs of locomotion for top-down control in traffic scenes. Two kinds of constraints could be derived: spatial constraints through knowledge about the applicability of motion concepts, and motion constraints through knowledge about typical motion. We proposed to compute motion constraints using a spatio-temporal buffer as a shared representation for bottom-up and top-down processes. Within the buffer motion concepts are expressed as typicality distributions from which predictions about object motion can be derived. A local prediction algorithm allows for the computation of search areas for low-level motion analysis. A low-level motion representation based on spatio-temporal Gabor cells is well suited for the integration of this kind of top-down information.

We presented an example where this procedure has been implemented. Using top-down guidance, the complexity of computation could be reduced significantly. Instead of analyzing the whole scene at the same level of detail, 1) only a small area could be chosen for an analysis and 2) the analysis could be focussed on specific spatio-temporal behavior within the area of interest.

5. References

[**Adelson + Bergen 85**]: Spatiotemporal energy models for the perception of motion, Journal of the Optical Society of America, A 2, 1985, pp. 284-299. [**Fleet 88**]: Implementation of Velocity-Tuned Filters and Image Encoding, Mitteilung, FBI-HH, Universität Hamburg, 1988. [**Kosslyn 80**]: Image and Mind, Harvard University Press, 1980. [**Mohnhaupt + Neumann 89**]: Some aspects of learning and reorganisation in an analogical representation, in 'Knowledge representation and organisation in machine learning', K. Morik (Ed.), Lecture Notes in Artificial Intelligence, Springer Verlag 1989, pp. 50-64. [**Mohnhaupt + Neumann 90**]: Understanding Object Motion: Recognition, Learning and Spatio-Temporal Reasoning, to appear in 'Journal of Robotics and Autonomous Systems', North Holland 1990. [**Neumann 89**]: Natural Language Description of Time-Varying Scenes, in 'Semantic Structures', David L. Waltz (Ed.), Lawrence Erlbaum, Hillsdale N.Y., 1989, pp. 167-207. [**Tsotsos 87**]: Image Understanding, in: S. Shapiro (Ed.), The encyclopedia of artificial intelligence, pp. 389-409, John Wiley and Sons, New York.

We thank David Fleet for many interesting comments and discussions.