High-Level Expectations for Low-Level Image Processing

Lothar Hotz[2], Bernd Neumann[1], and Kasim Terzic[1]

[1] Cognitive Systems Laboratory, Department Informatik, Universität Hamburg
22527 Hamburg, Germany
{terzic|neumann}@informatik.uni-hamburg.de
[2] HITeC e.V. c/o Department Informatik, Universität Hamburg
22527 Hamburg, Germany
hotz@informatik.uni-hamburg.de

Abstract. Scene interpretation systems are often conceived as extensions of low-level image analysis with bottom-up processing for high-level interpretations. In this contribution we show how a generic high-level interpretation system can generate hypotheses and initiate feedback in terms of top-down controlled low-level image analysis. Experimental results are reported about the recognition of structures in building facades.

1 Introduction

In recent years, growing interest in artificial cognitive systems has brought about increased efforts to extend the capabilities of computer vision systems towards higher-level interpretations [2, 13, 17, 15, 6, 5, 9]. Roughly, a high-level interpretation can be defined as an interpretation beyond the level of recognised objects. Typical examples are monitoring tasks (e.g. detecting a bank robbery), analysing traffic situations for a driver assistance system or interpreting aerial images of complex man-made structures. While existing approaches to high-level interpretation differ in many respects, they have in common that prior knowledge about spatial and temporal relations between several objects has to be brought to bear, be it in terms of probabilistic models [15], frame-based models called aggregates, logic-based conceptual descriptions [13], Situation Graph Trees [12] or Scenarios [4]. In the following, we will use the term "aggregate" for meaningful multiple-object units of a high-level scene interpretation.

In extending vision to high-level interpretations, one of the challenges is to exploit expectations derived from high-level structures for improved low-level processing. There exists much work addressing expectation-guided image analysis [12, 16, 11, 3], but to our knowledge few vision systems with a generic architecture have been proposed which allow to feed back expectations from aggregates at arbitrarily high levels of abstraction to image analysis procedures at the level of raw images. Nagel [1] has been one of the first to demonstrate with concrete experiments in the street traffic domain that high-level hypotheses about intended vehicle behaviour could in fact be used to influence the tracking unit and thus improve tracking under occlusion.

In this contribution we show how a scene interpretation system based on aggregates, introduced as generic high-level conceptual units in [8, 14, 13], can generate feedback in a generic manner. This is demonstrated by experiments with the fully implemented scene interpretation system SCENIC. One of the core mechanisms of SCENIC is the capability of part-whole reasoning, i.e. the capability of establishing an aggregate instantiation based on evidence for any of the aggregate parts. This allows to generate strong expectations about further evidence for parts of this aggregate and to feed back these expectations to lower levels. In SCENIC, this feedback process has been extended to control image-analysis processes below the level of recognised objects which provide the input to the high-level interpretation system.

In Section 2 we will describe the generic high-level reasoning facilities which may lead to object hypotheses supported only by the high-level context. The middle layer, called Match Box in SCENIC, is described in Section 3. It has the task to mediate between hypotheses and evidence, including the initiaton of goal-oriented low-level image analysis. Experiments and a summary are given in Section 4 and 5 respectively.

2 Top-Down Expectation Generation

In this section we describe techniques developed in SCENIC to generate high-level hypotheses about a scene, possibly containing incomplete information, and propagate consequences top-down to influence low-level processing. It is shown that this can be achieved by navigating in a highly structured interpretation space with a fixed set of interpretation steps. Expectations are generated by passing information obtained from evidence upwards and downwards along taxonomical and compositional hierarchies, and laterally to parts of the same aggregate.

High-level scene interpretation in SCENIC is based on conceptual knowledge about aggregates and their parts, embedded in compositional and taxonomical hierarchies (illustrated in Fig. 1). With `Scene` as the root of the compositional hierarchy, the conceptual knowledge base implicitly represents all possible scene interpretations. Each concept is described by attributes with value ranges or value sets, constraints between attributes and relations to other concepts. Spatial attributes representing the potential position and size of the scene objects in terms of ranges are of primary importance. They are represented as constraints which are automatically adjusted as new evidence or related hypotheses restrict the attribute values. Aggregate concepts have a generic structure:

aggregate name	contains a symbolic ID
parent concepts	contains IDs of taxonomical parents
external properties	provide a description of the aggregate as a whole
parts	describe the subunits out of which an aggregate is composed
constraints	specify which relatins must hold between the parts

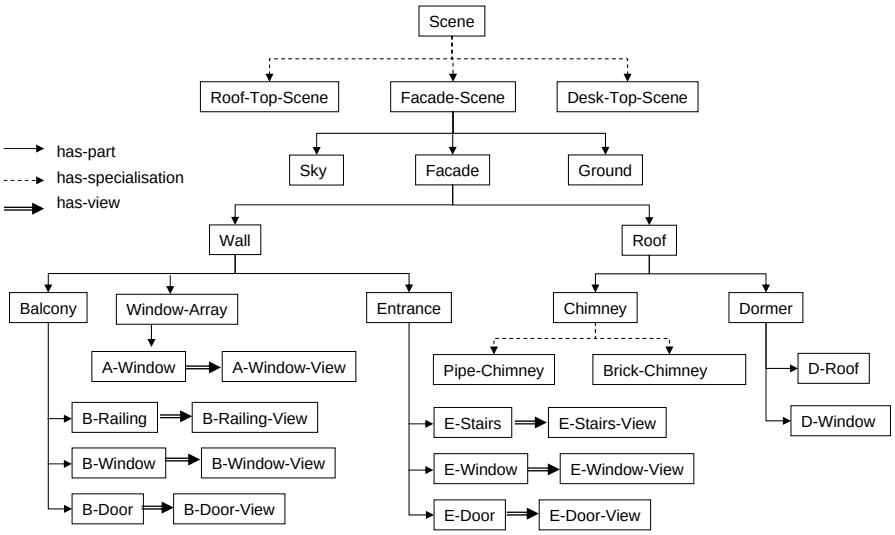

Fig. 1. Structure of facade knowledge base used for high-level interpretation in SCENIC Evidence can be related to object concepts by means of view concepts which specify properties of possible object views. It is the task of the middle layer (Section 3) to assign evidence to view concepts.

The conceptual knowledge base is completely described in terms of aggregate structures (including primitive objects which are aggregates without parts). Note that this does not allow constraints between parts of different aggregates. This restriction is intended to channel information flow exclusively along the structure of the compositional and taxonomical hierarchies.

The interpretation process obeys the following basic algorithm:

```
Repeat
Check for goal completion
Check for new evidence
Determine possible interpretation steps and update agenda
Select from agenda one of
{  evidence matching,
   aggregate instantiation,
   aggregate expansion,
   instance specialization,
   parameterization,
   constraint propagation }
Check for conflict
end
```

As elucidated in [8, 14, 13], the interpretation steps allow to construct all partial models of the knowledge base consistent with the evidence. Typically, we are interested in a single interpretation which meets a given goal, for example, to instantiate the concept `Scene` and its parts down to the primitive objects of

the compositional hierarchy such that the views of primitive objects optimally match evidence. Note that this allows to hypothesise objects without evidence - indispensable for realistic scene interpretation with occlusions, deficient low-level evidence, model limitations etc. Within the logic-based framework presented so far, there is no preference between consistent interpretations. A probabilistic framework providing a preference measure will be presented in a forthcoming publication.

We now describe the information flow, based on generic interpretation steps, which leads to top-down expectations and possible control of subsequent low-level operations. As an exemplary situation, we consider the recognition of a window-array based on evidence of a single window and causing expectations about more windows in the vicinity and at the same height.

Step 1 (evidence matching): Low-level evidence is assigned to `A-Window-View`, and a corresponding instance is created incorporating evidence properties (e.g. location and shape). This step is performed by the Match Box.

Step 2 (physical-object instantiation): The view instance leads to an instantiation of the corresponding physical-object concept `A-Window`. Image properties are transformed into physical-object properties using constraints between the physical-object concept and its views. This step is performed automatically and hence not subject to selection from the agenda as other interpretation steps.

Step 3 (aggregate instantiation): The `A-Window` is tentatively interpreted as part of the aggregate `Window-Array`. Constraints between external properties of the aggregate and its parts are set up and give rise to the spatial range where additional windows may be expected.

Step 4 (aggregate expansion): An additional `A-Window` instance is created as part of the `Window-Array`, meeting the constraints between the aggregate and its parts. While the size is closely restricted by the evidence for the first window, its location obeys loose restrictions specified in the conceptual description of the `Window-Array`. At this time, the instance is hypothesised by part-whole reasoning and not supported by evidence.

Step 5 (view instantiation): The window hypothesis leads to the instantiation of a corresponding view hypothesis with properties specified in image coordinates. In particular, the range of possible window locations is now expressed as an image area and the window size is given in pixel size. This step is performed automatically analog to Step 2.

Step 6 (evidence matching): The hypothesised view, restricted by constraints in Steps 4 and 5, is matched to evidence. This allows goal-oriented low-level image analysis meeting the restrictions of the view hypothesis. A decision has to be made whether the hypothesis has to be refuted because of missing or conflicting evidence.

In this report, we emphasise the use of feedback for low-level image analysis to resolve cases of insufficient evidence. But also other kinds of analysis have to be invoked such as occlusion reasoning, illumination and shadow analysis. All this is the task of the Match Box.

3 Middle Layer

The middle layer, called Match Box in SCENIC, acts as an interface between image analysis and symbolic interpretation. Its main tasks are:

- matching low-level evidence to high-level concepts,
- confirming or refuting high-level hypotheses, and
- initiating low-level activities from high-level hypotheses.

The combination of these tasks allows the interpretation system to operate in a feedback loop as originally proposed by [10] almost three decades ago.

Matching low-level evidence to high-level concepts is basically object classification. There are, however, several differences which complicate the task. First, the object classes for choice are not predefined but are influenced by restrictions resulting from the high-level context generated in preceding interpretation steps, typically including a ROI (region of interest) and size restrictions. Second, there is the problem of evidence assignment: Several pieces of evidence may qualify for the range restrictions of a hypothesis, and one has not only to choose between classes but also between evidences. The SCENIC Match Box currently operates in a simplified setting, where preclassified evidence is matched to views by selecting pairings with maximal spatial overlap.

Confirming and refuting high-level hypotheses is an additional aspect of the task. Hypotheses may be false due to bad high-level interpretation choices, hence it is necessary to compare theses hypotheses with relevant evidence. Currently, the Match Box distinguishes between two evidence qualities, "primary evidence" by low-level analysis with a high significance threshold, and "secondary evidence" with a lower significance threshold. The idea is to initially interpret the image based on primary evidence and invoke a refined image analysis to obtain secondary evidence only if required.

To compare a hypothesis with evidence, the Match Box has access to the view concepts of the knowledge base. If there is evidence within the ROI matching the evidence types of the hypothesis, the hypothesis can be confirmed. If there is conflicting evidence, it must be refuted. If there is no primary evidence, low-level image analysis is initiated to obtain secondary evidence. Here, the Match Box has a repertoire of image propcessing modules (IPMs) at its disposal with parameters which can be set corresponding to expectations.

4 Experiments

The experimental scene interpretation system SCENIC has been applied to numerous images of a database of ca. 600 building facades assembled in the EU-funded project eTRIMS (eTraining for the Interpretation of Man-made Scenes). The thrust of the project is to develop learning methods for structured objects, and the conceptual knowledge base of SCENIC actually comprises several learnt concepts, including `Window-Array`, `Balcony` and `Entrance` [7].

To demonstrate the feedback cycle, a rectified facade image has been processed by a low-level image analysis procedure trained to discover T-style windows. Fig. 2 shows the resulting primary evidence. As can be expected from bottom-up image analysis of a natural scene, the results also contain several false positives and false negatives.

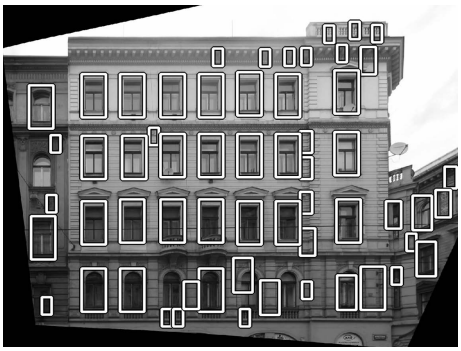

Fig. 2. Evidence created with an image processing module (IPM) trained to recognise windows.

The Match Box receives this evidence and creates window views from evidence items with high confidence value. It then passes these views to the interpretation system.

The interpretation system has been initiated to interpret a facade scene by instantiating the corresponding root node of the conceptual knowledge base. The initialisation also includes information relating image coordinates to world coordinates. The first interpretation phase has the goal to instantiate all obligatory descendants of the instantiated root node, in this case `Facade` and `Wall`.

In the second phase, the interpretation system creates physical-object instances corresponding to the view instances received from the Match Box. Furthermore, all aggregates which follow uniquely from the available instances are created in a bottom-up manner. In the experiment, instances of the concept `Window-Array` are created, when three windows exist which fulfil the `Window-Array` constraints. The concept `Window-array` specifies the following information:

```
Number of parts is [3 to inf].
All parts have type window.
All parts have similar y-position.
Any two neighbouring parts have similar distance.
All parts have similar height.
```

While the preceding steps have been obligatory, the next phase deals with uncertain interpretation decisions. The aggregates now trigger the creation of hypotheses for not yet instantiated optional parts. In the experiment, additional windows are searched at appropriate positions inferred from the established parts to complement the existing rudimentary window arrays. First, established parts

integrated, then new parts are hypothesised if there are remaining gaps. Fig. 3 (left) shows the resulting window arrays with three gaps filled by hypothesised windows.

Fig. 3. Left: Hypothesised window-arrays with four additional window hypotheses. Right: Result after restarting the low-level IPM with a hypothesised region of interest.

The interpretation system now creates view hypotheses for the newly hypothesised physical windows and asks the Match Box to confirm or refute these view hypotheses. As described in Section 3, the Match Box tries to match the views to existing evidence. In the experiment, there is insufficient evidence to confirm the hypotheses, so the Match Box initiates low-level image analysis of the ROIs with parameters set for weak evidence. The feedback results are shown in Fig. 3 (right): three new correct window evidences in regions where no window has been recognised in the first run (first three from left), and one new false window evidence in the wall area on the right. This confirms all window hypotheses generated by high-level interpretation.

5 Summary and Future Work

In this paper we have presented a generic approach for combining low-level image analysis and high-level interpretation so that feedback to low-level analysis can be achieved. To this end, a middle layer has been introduced which mediates between low-level evidence and high-level hypotheses. The combination of low-level, middle-layer and high-level techniques has been implemented in the system SCENIC, experimental results have been presented demonstrating the use of feedback for images with building facades. Feedback has been shown to allow a coarse-to-fine strategy, with fine-grained analysis only where required. Feedback also allows to invoke procedures specialised for the kinds of views which are to be verified. For example, texture analysis can be invoked for the verification of a wall background or an occluding tree.

In ongoing work, we integrate a probabilistic model for the compositional concept hierarchy to provide a preference measure for interpretation steps. Future

work will also deal with time-yarying scenarios, where high-level expectations often concern future events and are particularly valuable for scene interpretation.

References

1. M. Arens, A. Ottlik, and H.-H. Nagel. Using Behavioral Knowledge for Situated Prediction of Movements. In *Proc. 27th German Conference on Artificial Intelligence (KI-2004)*, volume LNAI 3238, pages 141–155. Springer, September 2004.
2. M. Borg, D. Thirde, J. Ferryman, F. Fusier, V. Valentin, F. Bremond, and M. Thonnat. A Real-Time Scene Understanding System for Airport Apron Monitoring. In *Proc. of IEEE International Conference on Computer Vision Systems ICVS06*, 2006.
3. E. Dickmanns. Expectation-based Dynamic Scene Understanding. In A. Blake and A. Yuille, editors, *Active Vision*, 1993.
4. B. Georis, M. Mazière, F. Brémond, and M. Thonnat. Evaluation and Knowledge Representation Formalisms to Improve Video Understanding. In *Proc. of IEEE International Conference on Computer Vision Systems ICVS06*, 2006.
5. R. Gerber and H.-H. Nagel. 'occurrence' Extraction from Image Sequences of Road Traffic Scenes. In L. van Gool and B. Schiele, editors, *Proceedings Workshop on Cognitive Vision*, pages 1–8. ETH Zurich, Switzerland, 2002.
6. S. Gong and H. Buxton. Understanding Visual Behaviour. *Image and Vision Computing*, 20(12):825–826, 2002.
7. J. Hartz and B. Neumann. Learning a knowledge base of ontological concepts for high-level scene interpretation. In *International Conference on Machine Learning and Applications*, Cincinnati (Ohio, USA), December 2007.
8. L. Hotz and B. Neumann. Scene Interpretation as a Configuration Task. *Künstliche Intelligenz*, 3:59–65, 2005.
9. Richard J. Howarth and Hilary Buxton. Conceptual descriptions from monitoring and watching image sequences. *Image and Vision Computing*, 18(2):105–135, 2000.
10. T. Kanade. Region Segmentation: Signal vs. Semantics. In *Proc. International Joint Conference on Pattern Recognition (IJCPR '78)*, Kyoto, Japan, 1978.
11. M. Mohnhaupt and B. Neumann. On the Use of Motion Concepts for Top-Down Control in Traffic Scenes. In *Proc. ECCV-90*, pages 598–600. Springer, 1990.
12. H.-H. Nagel. Natural Language Description of Image Sequences as a Form of Knowledge Representation. In *23. Fachtagung für Künstliche Intelligenz (KI99)*, pages 45–60. Springer, 1999.
13. B. Neumann and R. Möller. On Scene Interpretation with Description Logics. In *Cognitive Vision Systems*, volume LNCS 3948, pages 247–275. Springer, 2006.
14. B. Neumann and T. Weiss. Navigating through Logic-based Scene Models for High-level Scene Interpretations. In *3rd International Conference on Computer Vision Systems - ICVS 2003*, pages 212–222. Springer, 2003.
15. K. Sage, J. Howell, and H. Buxton. Recognition of Action, Activity and Behaviour in the ActIPret Project. *Künstliche Intelligenz*, 3:30–33, 2005.
16. J.M. Tenenbaum and H.G. Barrow. Experiments in Interpretation Guided Segmentation,. *Artificial Intelligence Journal*, 8(3):241–274, 1977.
17. M. Vincze, W. Ponweiser, and M. Zillich. Contextual Coordination in a Cognitive Vision System for Symbolic Activity Interpretation. In *Proc. of IEEE International Conference on Computer Vision Systems ICVS06*, 2006.