ON THE LOGICS OF IMAGE INTERPRETATION: MODEL-CONSTRUCTION IN A FORMAL KNOWLEDGE-REPRESENTATION FRAMEWORK

Carsten Schröder and Bernd Neumann

Universität Hamburg, Fachbereich Informatik,
Vogt-Kölln-Straße 30, 22527 Hamburg, Germany,
{schroeder,neumann}@informatik.uni-hamburg.de

ABSTRACT

In this contribution, we present a formal, logic-based approach to image interpretation by combining methods from two different research areas, namely computer vision and knowledge representation. After describing two well known approaches we presented a concise definition of the required solution of an image understanding problem. We then propose an object-centered, KL-ONE-like description logic tailored to the representation needs in image understanding and a calculus for computing an interpretation of a given image according to our definition.

1. INTRODUCTION

Informally, *image interpretation* is the task of creating a description of a scene, i.e. a part of our world, depicted in an image or sequence of images. The result includes a description of objects existing in the scene and properties of the scene itself which are not bound to any specific objects. If image sequences are analyzed instead of single images, image interpretation includes a description of occurrences, events, processes, episodes, histories or even intentions as well, depending on the length of the sequence.

In order to be useful, these descriptions must be given by using the notions we have about our world. What is needed, therefore, is a modeling of the different classes of objects and types of events clearly defining the properties of their possible instances in a scene. Knowledge representation is the field of AI which deals with suitable formalisms which can be used for defining concepts, i.e. it provides languages, and for reasoning about instances of these concepts, i.e. it provides inference mechanisms. While a representation language is definitely needed for our task, it is not quite clear, however, what kind of inference mechanisms are needed. Is image interpretation a deductive reasoning process, or is abductive, plausible, or even probabilistic reasoning necessary?

What we are interested in is a formal definition of image interpretation which, first, helps us to a better understanding of the problem, second, enables us to develop appropriate solutions, and third, lays the ground for standardized and reusable application software. In order not to search for methods in vain, let us first describe a particular problem we have in mind: Imagine a surveillance task where in regular, though quite long, intervals aerial images are taken from geographic areas with known coordinates (the images are assumed to be registered), and assume these areas are known, i.e. we have some sort of reference information describing the state of the areas at some previous time. We would like to know then the new state and the changes which have taken place in the meantime. The descriptions of changes should be given on a rather high semantic level, e.g. something like "the runway of the airport was elongated" would be appropriate.

In Section 2 we analyze two well-known, logic-based approaches from the literature. In Section 3 we present our definition of the problem, which builds on the previous two, then describe a representation language, and finally sketch a calculus which can be used for computing an image interpretation given the knowledge about the contents of an image and a definition of the relevant concepts. The paper ends with a conclusion.

2. RELATED WORK

The first concise definition of the problem of image interpretation—in a logical framework—was given by Reiter & Mackworth in their reconstruction of the MAPSEE-approach to sketch map understanding [1]. They showed how the relevant generic knowledge about the domain and the *a priori* given contents of a sketch map (in a symbolic form in terms of lines and regions) can be represented using first order predicate logic. Then they made use of the model-theoretic semantics of first order logic and defined an interpretation of a sketch map as a logical *model* of the set of first order formulas expressing the given knowledge, while a model, to recall this, is a logical interpretation satisfying a given set of formulas.

Taking this definition, the answer to our question is as follows: Image interpretation is not simply a logical inference process; neither a deductive one, which only generates formulas valid under *all* models of the given set of formulas, nor an abductive one, which generates explanations. It is the (re-)construction of a specific possible world[1] consistent with the given knowledge.

As Reiter & Mackworth already pointed out, the problem with their approach is that, in general, first order predicate logic is undecidable. Therefore, it cannot be checked

[1] In contrast to the notions of *logical models* and *interpretations* we use this term in its non-technical sense here.

0-7803-3258-X/96/$5.00 © 1996 IEEE

whether a given set of formulas has a model at all, and if so, the models may not be finite. Fortunately, a scene depicted in an image is always finite (on a limited scale, at least), which ensures the finite model property. Reiter & Mackworth make use of this by including a domain closure axiom which enumerates all the objects in a sketch map to be interpreted as well as a unique name axiom. This enables them to transform the given set of formulas to propositional logic. The constraint solving techniques developed in the MAPSEE-project can be used then for model construction. So, using this definition, image interpretation can also be seen as a constraint satisfaction problem. However, while the finite nature of scenes allows to construct the required models, it is still not possible to check the consistency of the scene-independent generic domain knowledge, for which a domain closure axiom cannot be specified. In addition, it seems that in [1] the process of transforming the set of formulas to propositional form was performed by hand.

Motivated by the work of Reiter & Mackworth, a second definition of image interpretation was given by Matsuyama & Hwang as a result of a logical reconstruction of their SIGMA-system for aerial image understanding [2]. In this reconstruction, they were using first order logic as the representation language as well, but in contrast to Reiter & Mackworth they did not require the contents of the images, i.e. a segmentation, to be completely given *a priori*. They see image interpretation as an instance of the *hypothesis-based reasoning* approach proposed by Poole et al. [3], whose task is to find a set of formulas, the *hypotheses* or *explanations*, which is consistent with the given knowledge and complements the generic domain knowledge, so that the formulas representing the *observations*, i.e. the contents of the image, can be logically deduced. As a segmentation of the image is not completely given *a priori*, checking consistency requires to further analyze the image whenever any new hypotheses concerning its contents are generated. This amounts to an incremental, expectation-driven image analysis.

Taking this definition, image interpretation is an abductive reasoning process generating explanations. However, in contrast to the model construction approach of Reiter & Mackworth, it does not make all the information explicit which is implied by the given knowledge and the generated hypotheses, for it is not required to perform any deductive inferences at all. In addition, it can easily be shown that in some cases of incomplete knowledge the explanations to be generated must be equal to the observations themselves, something we would hesitate to call an explanation. Looking at the definition in another way, image interpretation can be seen as the construction of partial logical models which are not propositionally closed.

3. THE APPROACH

In the following we present a formal approach to image interpretation which is similar to those just described in its use of a logical representation language having a clearly defined model-theoretic semantics, but (1) is not based on an *a priori* given segmentation, (2) provides a language suitable for expressing the geometric properties of objects, and (3) provides an effective calculus not relying on any hand transformation. We start by giving a definition which is suited to the surveillance task described in the introduction.

3.1. The Definition

In comparison of the two approaches described above, the definition of the problem given by Reiter & Mackworth seems to be the more natural one. However, the assumption that image interpretation can satisfactorily be done based on an *a priori* given segmentation of an image was proven to be quite unreasonable in recent years. Therefore, we are going to drop this assumption in the following. In addition, we will be satisfied with partial models not spelling out every single detail of a possible world.

Definition 1 (Interpretation of an Image) *Let $\mathcal{F}$ be a set of axioms of a classical logical language $\mathcal{L}$ containing the scene-independent generic domain knowledge as well as the given reference about the scene depicted in the image I, and let $\mathcal{B}$ be a subset of the axioms of $\mathcal{L}$ representing the observations which can potentially be extracted from the image I. An* interpretation *of the image I is then defined to be a partial logical model $\mathcal{I}_p = \langle \mathcal{D}_p, \cdot^{\mathcal{I}_p} \rangle$ of the set of axioms $\mathcal{F} \cup \mathcal{B}$ (with $\mathcal{D}_p$ as the domain of discourse and $\cdot^{\mathcal{I}_p}$ as a partial interpretation function) satisfying the following conditions:*

Consistency: *$\mathcal{I}_p$ can be extended to a complete model $\mathcal{I} = \langle \mathcal{D}, \cdot^{\mathcal{I}} \rangle$ with $\mathcal{D} \supseteq \mathcal{D}_p$ and $\cdot^{\mathcal{I}}$ being an extension of $\cdot^{\mathcal{I}_p}$.*

Specialty: *If $\mathcal{F} \cup \mathcal{B} \models_{\mathcal{I}_p} A \vee B$, but not $\models A \vee B$, then $\mathcal{F} \cup \mathcal{B} \models_{\mathcal{I}_p} A$ or $\mathcal{F} \cup \mathcal{B} \models_{\mathcal{I}_p} B$, while A and B are propositions, i.e. variable free formulas of $\mathcal{L}$.*

The Specialty condition specifies how complete the solution is required to be. Firstly, it ensures that the partial model is deductively closed. As the propositional deductive closure is much to small to be an interesting result, the Specialty condition, secondly, enforces the resolution of non-tautological disjunctive propositions in order to get a more specific image interpretation. If we would drop the exception that tautological disjunctive propositions do not have to be resolved, then a solution would be required to be a complete model.

This definition allows for incrementally acquiring the observations. Note, however, that we might end up with a result not containing any new information about the scene at all. This happens whenever we start with an empty set $\mathcal{B}$ and the set $\mathcal{F}$ does not imply the existence of any objects in the new image.

3.2. The Language

We are looking for a language which, first, is more expressive than plain predicate logic in order to be useful for interpreting real images. In particular, the geometrical properties of objects like absolute as well as relative position, orientation, shape, and size should be expressible, for this is the most important kind of knowledge we have. Second, we would like that at least a fragment of the language, which is needed for expressing the scene-independent knowledge,

$$
\begin{array}{rcll}
C, D & \rightarrow & CN \mid & \\
 & & \top \mid & \mathcal{D} \\
 & & \bot \mid & \emptyset \\
 & & C \sqcap D \mid & C^{\mathcal{I}} \cap D^{\mathcal{I}} \\
 & & C \sqcup D \mid & C^{\mathcal{I}} \cup D^{\mathcal{I}} \\
 & & \neg C \mid & \mathcal{D} \setminus C^{\mathcal{I}} \\
 & & (\geq n\, \varrho) \mid & \{a \in \mathcal{D} \mid \|\{b \in \mathcal{D} \mid (a,b) \in \varrho^{\mathcal{I}}\}\| \geq n\} \\
 & & (\leq n\, \varrho) \mid & \{a \in \mathcal{D} \mid \|\{b \in \mathcal{D} \mid (a,b) \in \varrho^{\mathcal{I}}\}\| \leq n\} \\
 & & \forall\,(\varrho_1 \ldots \varrho_n)\,.\,\Gamma \mid & \{a \in \mathcal{D} \mid \forall b_1, \ldots, b_n\colon ((a,b_1) \in \varrho_1^{\mathcal{I}} \wedge \ldots \wedge (a,b_n) \in \varrho_n^{\mathcal{I}}) \Rightarrow ((b_1, \ldots, b_n) \in \Gamma^{\mathcal{I}})\} \\
 & & \exists\,(\varrho_1 \ldots \varrho_n)\,.\,\Gamma & \{a \in \mathcal{D} \mid \exists b_1, \ldots, b_n\colon (a,b_1) \in \varrho_1^{\mathcal{I}} \wedge \ldots \wedge (a,b_n) \in \varrho_n^{\mathcal{I}} \wedge (b_1, \ldots, b_n) \in \Gamma^{\mathcal{I}}\} \\
R, S & \rightarrow & RN \mid & \\
 & & R \sqcap S \mid & R^{\mathcal{I}} \cap S^{\mathcal{I}} \\
 & & R \sqcap (C \times D) \mid & R^{\mathcal{I}} \cap (C^{\mathcal{I}} \times D^{\mathcal{I}}) \\
 & & R^{-1} & \{(b,a) \mid (a,b) \in R^{\mathcal{I}}\} \\
\varrho & \rightarrow & R_1 \circ \ldots \circ R_n & R_1^{\mathcal{I}} \circ \ldots \circ R_n^{\mathcal{I}} \\
\Gamma & \rightarrow & C \mid \text{ if } n = 1 & \\
 & & \Pi \quad \text{if } n \geq 1 & \\
\Pi & \rightarrow & P \mid & \\
 & & \forall\,(\varrho_{11} \ldots \varrho_{1m_1}) \ldots (\varrho_{n1} \ldots \varrho_{nm_n})\,.\,\Pi \mid & \text{semantics analogous to } (\forall\,(\varrho_1 \ldots \varrho_n)\,.\,\Gamma)^{\mathcal{I}} \\
 & & \exists\,(\varrho_{11} \ldots \varrho_{1m_1}) \ldots (\varrho_{n1} \ldots \varrho_{nm_n})\,.\,\Pi & \text{semantics analogous to } (\exists\,(\varrho_1 \ldots \varrho_n)\,.\,\Gamma)^{\mathcal{I}} \\
P & \rightarrow & \text{any } n\text{-ary first order predicate with numeric (possibly non-linear) inequalities as atomic formulas} &
\end{array}
$$

Figure 1: Syntax and semantics of the terminological language.

$CN \doteq C$	$CN \sqsubseteq C$	$(\geq n\, C)$	$(\leq n\, C)$	$a\colon C$	$(a,b)\colon \varrho$	$(a_1, \ldots, a_n)\colon P$	$(a,b) :=$	$(a,b) :\neq$
$CN^{\mathcal{I}} = C^{\mathcal{I}}$	$CN^{\mathcal{I}} \subseteq C^{\mathcal{I}}$	$\|C^{\mathcal{I}}\| \geq n$	$\|C^{\mathcal{I}}\| \leq n$	$a^{\mathcal{I}} \in C^{\mathcal{I}}$	$(a^{\mathcal{I}}, b^{\mathcal{I}}) \in \varrho^{\mathcal{I}}$	$(a_1^{\mathcal{I}}, \ldots, a_n^{\mathcal{I}}) \in P^{\mathcal{I}}$	$a^{\mathcal{I}} = b^{\mathcal{I}}$	$a^{\mathcal{I}} \neq b^{\mathcal{I}}$

Figure 2: Syntax and semantics of the assertional language.

is decidable. We would be able to check this knowledge for consistency then.

We propose to use an object-centered KL-ONE-like *description logic* [4] which provides a language for describing *concepts* (denoted by C and D in the following) by using *concept names* and *role names* (denoted by CN and RN). Figure 1 shows the syntax and semantics of our concept language. It is the result of a careful analysis of the required expressivity and the aim of keeping at least a certain fragment of it decidable [5].

A logical interpretation $\mathcal{I} = \langle \mathcal{D}, \cdot^{\mathcal{I}} \rangle$ of our language maps every concept name, role name, object name (denoted by a and b and used below), and every n-ary numerical predicate P, respectively, as follows: $CN^{\mathcal{I}} \subseteq \mathcal{D}$, $RN^{\mathcal{I}} \subseteq (\mathcal{D} \cup \mathbb{R})^2$, $a^{\mathcal{I}} \in \mathcal{D} \cup \mathbb{R}$, and $P^{\mathcal{I}} \subseteq \mathbb{R}^n$. The semantics of the various terms of the language is shown in the second column of Figure 1.

A knowledge base of our formalism consists of two components, an *intensional* or *terminological* one, called TBox, and an *extensional* or *assertional* one, called ABox. They may contain terminological and assertional axioms, respectively, as shown in the first line of Figure 2. The first four are terminological axioms; two for expressing universal knowledge by defining concepts in terms of necessary and maybe even sufficient conditions, two for expressing contingent knowledge by specifying constraints on the minimum and maximum number of instances of a given concept. The final five are assertional axioms used for stating facts about objects. The second line of Figure 2 specifies the semantics of the axioms by stating when an axiom is satisfied by a logical interpretation of our language.

In the following we show two concept definitions as an example. Suppose we are in an airport domain and want to define the concept of a runway. One of its most prominent properties (besides its shape) is that there must be a taxiway parallel to it in a certain distance and connected to it by at least 2 driveways. The following terminological axiom tries to formalize this by expressing at least some of these conditions:

Runwaylike-Object $\doteq$

Roadobject $\sqcap$ ($\geq$ 2 has-connecting-driveway) $\sqcap$

$\exists$ (frame has-connecting-driveway-neighbor) .

$\exists$ ((direction $\circ$ x) (direction $\circ$ y))

((frame $\circ$ direction $\circ$ x) (frame $\circ$ direction $\circ$ y)) . parallep

The concept term on the right hand side describes the set of objects which are instances of the concept Roadobject,

have at least two fillers of the role has-connecting-driveway, and, in addition, satisfy the following conditions: (1) They have a filler of the role frame (supposed to be a coordinate frame) which in turn is required to have fillers of the x and y component of its direction vector. (2) They have a filler of the role has-connecting-driveway-neighbor (supposed to be the neighbor of a neighbor) which in turn is required to have fillers of the x and y component of the direction vector of its coordinate frame. (3) The components of the two direction vectors fulfill a 4-place numeric predicate expressing the constraint of being parallel (we omit specifying the predicate and use a name instead).[2] The set of objects so described is required to be equal to the set of instances of the concept name Runwaylike-Object, so the conditions are necessary as well as *sufficient*.

The language can be use to define possible types of changes as well. Changes between two states can be reified and represented by a change "object" which is connected to two spatial objects describing the former and the later state. Consider the elongation of an object. This is defined by a constraint on the length of an object in two different states:

$$\begin{aligned}\mathsf{Elongation} \doteq\ & \mathsf{Change\text{-}Object} \sqcap \\ & \exists\,((\mathsf{pre} \circ \mathsf{shape} \circ \mathsf{length}) \\ & \quad (\mathsf{post} \circ \mathsf{shape} \circ \mathsf{length}))\,.\,\lambda l_1, l_2.\,(l_1 < l_2)\end{aligned}$$

To summarize, our language can be used to formalize the scene-independent domain knowledge by using terminological axioms, as shown above. The reference information can be formalized by using assertional axioms, and the observations can be specified by using assertional as well as cardinality axioms. We conclude this section by showing how to express a domain closure by a very simple axiom: $(\leq n\,\top)$. The upper bound n for the number of objects can be estimated from the size and resolution of the image and the given reference.

3.3. The Calculus

Schmidt-Schauß & Smolka [6] presented a complete and sound consistency test for a subset of our language. It is based on the idea of a tableau calculus which tries to construct a model for a given knowledge base. This is done by completing the ABox of the knowledge base so that it becomes a model of itself as well as the TBox. We have extended this calculus to work for our language and equipped it with a method for performing an expectation-driven analysis of the image to be interpreted for any hypotheses about its contents which may have been generated. No transformation to propositional form is required by this calculus. It is guaranteed to be complete in the sense that all inconsistencies are detected, but as our language defined above is undecidable, it is only guaranteed to terminate if an axiom of the form $(\leq n\,\top)$ is included in the knowledge base. The calculus effectively computes one or all models of a given knowledge-base and image. Unfortunately, it turns out that the problem is highly intractable: it can be shown that it is not even in PSPACE. In addition, this calculus can be used for checking the consistency of the scene-independent domain knowledge using a fragment of our terminological language which is claimed (though not yet proven) to be decidable. Note, that a special reasoner is needed as part of our calculus for checking the consistency of a set of numeric predicates based on possibly non-linear (in-)equalities. This can be provided, however, by using the cylindrical algebraic decomposition method (CAD) proposed by Collins [7]. See [5] for the details of the calculus as well as a comprehensive description of the overall approach.

[2] Actually, the role names not starting with has- must be interpreted as partial functions here, in order for this description to be correct. This can easily be achieved by a slight though spacious change of the language.

4. CONCLUSION

We have discussed the problem of image interpretation in a formal, logical framework. After describing two well known views on this problem we presented our own approach: A concise definition of the problem, an object-centered description logic suited to the representation needs in image understanding, and an effective calculus. As the problem is intractable, a next step would be to analyze incomplete methods.

5. REFERENCES

[1] Raymond Reiter and Alan K. Mackworth. A logical framework for depiction and image interpretation. *Artificial Intelligence*, 41:125–155, 1989/90.

[2] Takashi Matsuyama and Vincent Shang-Shouq Hwang. *SIGMA: A Knowledge-Based Aerial Image Understanding System.* Plenum Press, New York – London, 1990.

[3] David Poole, Randy Goebel, and Romas Aleliunas. Theorist: A logical reasoning system for defaults and diagnosis. In Nick Cercone and Gordon McCalla, editors, *The Knowledge Frontier – Essays in the Representation of Knowledge*, chapter 13, pages 331–352. Springer-Verlag, Berlin – Heidelberg – New York, 1987.

[4] William A. Woods and James G. Schmolze. The KL-ONE family. In Fritz Lehmann, editor, *Semantic Networks in Artificial Intelligence*, pages 133–177. Pergamon Press, Oxford, 1992.

[5] Carsten Schröder. *Bildinterpretation durch Modellkonstruktion: Ein logikbasierter Ansatz zur rechnergestützten Analyse von Bildern.* Dissertation, Fachbereich Informatik, Universität Hamburg, 1996. In preparation.

[6] Manfred Schmidt-Schauß and Gert Smolka. Attributive concept descriptions with complements. *Artificial Intelligence*, 48(1):1–26, 1991.

[7] George E. Collins. Quantifier elimination for real closed fields by cylindrical algebraic decomposition. In H. Brakhage, editor, *Automata Theory and Formal Languages, 2nd GI Conference*, volume 33 of *Lecture Notes in Computer Science*, pages 134–183, Kaiserslautern, May 20–23, 1975. Springer-Verlag, Berlin – Heidelberg – New York, 1975.