Georgios Paliouras, Constantine D. Spyropoulos,
George Tsatsaronis

Knowledge-Driven Multimedia Information Extraction and Ontology Evolution

Volume 6050

Logical Formalization of Multimedia Interpretation

by Sofia Espinosa, Atila Kaya, Ralf Möller

August 23, 2010

Springer

Contents

1

Logical Formalization of Multimedia Interpretation

Nowadays, many documents in local repositories as well as in resources on the web are multimedia documents that contain not only textual but also visual and auditory information. Despite this fact, retrieval techniques that rely only on information from textual sources are still widely used due to the success of current text indexing technology. However, to increase precision and recall of multimedia retrieval, the exploitation of information from all modalities is indispensable, and high-level descriptions of multimedia content are required. These symbolic descriptions, also called deep-level semantic annotations, play a crucial role in facilitating expressive multimedia retrieval. Even for text-based retrieval systems, deep-level descriptions of content are useful (see, e.g., [BCSW07]).

There is a general consensus that manual annotation of multimedia documents is a tedious and expensive task which must be automated in order to obtain annotations for large document repositories. *Multimedia interpretation* is defined here as the process of producing deep-level semantic annotations based on low-level media analysis processes and domain-specific conceptual data models with formal, logical semantics.

The primary goal of this chapter is to present logical formalizations of interpretation. The chapter presents pioneering work on logic-based scene interpretation that has a strong influence on multimedia interpretation. Early approaches are discussed in more detail to analyze the main reasoning techniques. More recent approaches, which are more formal and therefore harder to understand, are referred to by providing references to the literature such that the reader can get an overview over the research field of logic-based media interpretation.

The discussion about scene interpretation is complemented with a presentation of logical approaches to text interpretation. Logical representations for deep-level video interpretation are discussed afterwards. The main goal of the chapter is to investigate the role of logic in the interpretation process. In order to focus on this goal, we neglect probabilistic approaches to this topic (but we give some pointers to the literature).

Logic-based media interpretation builds on initial symbolic descriptions of media content. In the next section, we argue that it is reasonable to expect so-called *multimedia analysis* processes to be able to reliably produce description about information that is, more or less, directly observable.

1.1 Prerequisites for Interpretation: Media Analysis

The identification of directly observable information in different modalities, also called *surface-level information*, has been studied in the past for at least three decades. In natural language processing, *information extraction* is one of the major tasks that aims to automatically extract structured information such as named entities and certain relations between entities. Evaluations have shown that state-of-the-art information extraction systems are very powerful language analysis tools that can recognize names and noun groups with an accuracy higher than 90% [CY99]. Different systems exploit various machine-learning techniques such as k-nearest neighbors or Hidden Markov Models. They have been successfully used for solving real-world problems [AHB$^+$93]. However, information extraction is a more restricted problem than general language understanding, and language analysis techniques employed in these systems provides for simple, reliable symbolic content descriptions but are not as powerful as full syntactic language analysis. A state of the art system for text analysis is OpenCalais (`http://www.opencalais.com`), which returns its results as annotations to a text in a logic-based language. However, when it comes to extracting more abstract information such as events that require a deep understanding of the domain, information extraction systems are reported not to perform well in general [Gri03].

In computer vision, *object recognition* aims to find objects in images (scenes) or image sequences (videos). Even though object recognition has been successfully applied in specific domains, e.g., for finding faces in images [VJ01], general object recognition is still an unsolved problem. In many approaches, object recognition follows segmentation, where images are partitioned into homogeneous regions, i.e. sets of pixels. The pixels in a region are similar w.r.t. some feature such as color, intensity or texture [SHB07]. The derivation of homogeneous regions is supported by techniques such as color histograms or shape analysis. However, when used without further knowledge resources, these “global” techniques are not appropriate for general-purpose object recognition in images [JB08]. Therefore, a wide range of local descriptors, such as Harris corners [HS88], Shape Context [BMP02] and Scale Invariant Transform (SIFT) [Low04], have been proposed. Nowadays, local descriptors are successfully used for solving practical problems. For example, SIFT has been applied to the problem of robot localization in unknown environments in robotics [SLL02]. Mikolajczyk and Schmid present a comprehensive evaluation of various local descriptors [MS05]. We would like to point out that logic-based representations have also been used at the analysis level (maybe in combination with probabilistic or fuzzy representations such as, e.g., in [SS08]).

Recently, Leibe and Schiele presented an approach that considers object recognition and segmentation as intertwined processes and uses top-down knowledge for guiding the segmentation process [LS03]. The authors reported on experimental results that show the capacity of the approach to categorize and segment diverse categories such as cars and cows. As a result, even though object and event recognition in the general domain is beyond the capabilities of current technology [KLSG03], the identification of observable information in image and video sequences in specific domains can indeed be achieved with state-of-the-art computer vision systems. Information extraction from text and the field of computer vision are related research fields providing the input required for the interpretation process.

Thus we can reasonably assume that the above-mentioned analysis processes can compute symbolic descriptions of media content, and make such descriptions available as input to multimedia interpretation processes. It is also very well possible that media analysis can be influenced by media interpretation. But for the time being we consider analysis and interpretation as sequential steps. In any case, the discussion reveals that recent advances in media analysis provide for a solid foundation to the derivation of deep-level abstract content descriptions based on a logical representation language.

1.2 Logic-based Scene Interpretation

In this section we present related work on scene interpretation that has a strong influence on the design of multimedia interpretation processes. In fact, the multimedia interpretation problem, for which also modalities beyond images are relevant, can be considered as a generalization of scene interpretation. Although there exist a substantial number of approaches to high-level scene interpretation in the literature, unfortunately, many of them are not built on representation languages with a formal semantics. In this section we focus on approaches that exploit formal, declarative representations for scene interpretation and that have been implemented as software systems. Our goal is not only to cite relevant work on scene interpretation but also to identify key problems in scene interpretation. We expect the reader to be familiar with first-order logic and, to some extent, with logic programming as well as description logic (see pointers to the literature in the text).

1.2.1 Scene Interpretation Based on Model Construction

The first formal theory of scene interpretation based on logics was introduced by Reiter and Mackworth [RM87]. They propose a so-called theory of depiction and interpretation that formalizes image-domain knowledge, scene-domain knowledge and a mapping between the image and scene domains using first-order logic [RM90]. An interpretation of an image is then defined as a logical model of a set of logical formulae which formalize background knowledge as well as the output of low-level scene analysis processes.

We shortly discuss the main ideas of the approach in [RM90], and we recapitulate the system *Mapsee*, which has been implemented for the interpretation of hand-drawn sketch maps of geographical regions [MMH88]. Given a sketch map consisting of chains[1], regions and various relations between them, the goal of the system is to compute an interpretation in terms of roads, rivers, shores, areas of land, and areas of water, etc.

The image-domain knowledge includes general knowledge about maps such as the taxonomy of image-domain objects, which are specified through first-order logic axioms:

$$\begin{array}{l}\forall x:\ \textit{image-object}(x)\ \Leftrightarrow\ \textit{chain}(x) \vee \textit{region}(x)\\ \forall x:\ \neg(\textit{chain}(x) \wedge \textit{region}(x))\end{array}$$

The first axiom states that chains and regions, so-called image primitives, are the only objects that can exist in a map, whereas the latter axiom states that an object cannot be both chain and region at the same time (disjointness of image primitives). Relations between image-domain objects are also part of the image-domain knowledge and are specified using predicates such as $\textit{tee}(c, c')$, $\textit{bound}(c, r)$ etc. For example, the predicate $\textit{tee}(c, c')$ means that chain c meets chain c$'$ at a T-junction.

The approach assumes a map description to consist of finitely many chains and regions together with finitely many relations between the chains and regions. Therefore, the system makes the *domain closure assumption* by postulating that all map objects are completely known. To this end, closure axioms of the following form are used (i_m and i'_n are constants):

$$\begin{array}{l}\forall x:\ \textit{chain}(x)\ \Leftrightarrow\ x\ =\ i_1 \vee \cdots \vee x\ = i_m\\ \forall x:\ \textit{region}(x)\ \Leftrightarrow\ x\ =\ i'_1 \vee \cdots\ \ \vee x\ = i'_n\\ \forall x, y:\ \textit{tee}(x, y)\ \Leftrightarrow\ (x\ =\ i_1 \wedge y\ =\ i'_1) \vee \cdots \vee\ (x\ =\ i_k \wedge y\ =\ i'_k)\\ \ldots\end{array}$$

Furthermore, the system makes the *unique name assumption* by assuming that all constants (e.g., image primitives such as chains and regions) denote different objects. Scene-domain knowledge is represented by axioms for objects such as roads, rivers, shores, or land and water areas. For instance, the following equivalence, coverage and disjointness axioms are used.

$$\begin{array}{l}\forall x:\ \textit{scene-object}(x)\ \Leftrightarrow\ \textit{linear-scene-object}(x) \vee \textit{area}(x)\\ \forall x:\ \textit{linear-scene-object}(x)\ \Leftrightarrow\ \textit{road}(x) \vee \textit{river}(x), \vee \textit{shore}(x)\\ \forall x:\ \neg(\textit{road}(x) \wedge \textit{river}(x))\\ \forall x:\ \neg(\textit{linear-scene-object}(x) \wedge \textit{area}(x))\ \ldots\end{array}$$

[1] Chain is the term used in the original paper for polylines.

In addition, the scene-domain knowledge base contains also specific restrictions such as, for instance, rivers do not cross each other:

$$\forall x, y : \ \mathit{river}(x) \wedge \mathit{river}(y) \Rightarrow \ \neg\ \mathit{cross}(x, y)$$

Axioms that restrict the domain and range of relations to scene objects only are also used:

$$\forall x, y : \ \mathit{cross}(x, y) \Rightarrow \ \mathit{scene\text{-}object}(x) \ \wedge \ \mathit{scene\text{-}object}(y)$$

Besides the specification of intra image- and scene-domain axioms, also inter-domain axioms between the image and scene domain are specified (so called mapping axioms). The mapping axioms are represented using the binary predicate $\mathit{\Delta}(i, s)$ meaning that image object i depicts scene object s. The depiction relation only holds between image and scene objects:

$$\forall i, s : \ \mathit{\Delta}(i, s) \Rightarrow \mathit{image\text{-}object}(i) \wedge \mathit{scene\text{-}object}(s)$$

For specifying image-scene-domain mappings, closure and disjointness axioms are provided.

$$\forall x : \ \mathit{image\text{-}object}(x) \vee \mathit{scene\text{-}object}(x)$$
$$\forall x : \ \neg(\mathit{image\text{-}object}(x) \wedge \mathit{scene\text{-}object}(x))$$

Furthermore, it is assumed that every image object i depicts a unique scene object, which is denoted by $\sigma(i)$:

$$\forall i : \ \mathit{image\text{-}object}(i) \Rightarrow \mathit{scene\text{-}object}(\sigma(i)) \wedge \mathit{\Delta}(i, \sigma(i)) \wedge [\forall s : \mathit{\Delta}(i, s) \Rightarrow s = \sigma(i)]$$

and every scene object is depicted by a unique image object:

$$\forall s : \ \mathit{scene\text{-}object}(s) \Rightarrow (\exists^1_i : \ \mathit{image\text{-}object}(i) \wedge \mathit{\Delta}(i, s))$$

The notation $\exists^1_i : \alpha(x)$ means that there exists exactly one x for which $\alpha(x)$ holds. Finally, mappings between the image- and scene-objects

$$\forall i, s : \ \mathit{\Delta}(i, s) \wedge \mathit{region}(i) \Rightarrow \mathit{area}(s)$$
$$\forall i, s : \ \mathit{\Delta}(i, s) \wedge \mathit{chain}(i) \Rightarrow \mathit{linear\text{-}scene\text{-}object}(s)$$

and mappings between relations of the image and scene domains are specified:

$$\forall i_1, i_2, s_1, s_2 : \ \mathit{\Delta}(i_1, s_1) \wedge \mathit{\Delta}(i_2, s_2) \Rightarrow \mathit{tee}(i_1, i_2) \Leftrightarrow \mathit{joins}(s_1, s_2)$$
$$\forall i_1, i_2, s_1, s_2 : \ \mathit{\Delta}(i_1, s_1) \wedge \mathit{\Delta}(i_2, s_2) \Rightarrow \mathit{chi}(i_1, i_2) \Leftrightarrow \mathit{cross}(s_1, s_2)$$
$$\ldots$$

The above-mentioned axioms state that *tee*[2] relations in the image depict *joins* relations in the scene and vice versa, whereas *chi*[3] relations in the image depict *cross* relations in the scene.

[2] Shorthand for T-junction.

[3] Shorthand for X-junction.

Given the specification of all relevant image-domain axioms, scene-domain axioms and mapping axioms, Reiter and Mackworth define an *interpretation* of an image, specified as set of logical facts, as a logical model of these facts w.r.t. the axioms in the knowledge base.

The main problem here is that, in principle, a set of first-order formulas may have infinitely many models, which in turn might be infinite, and, therefore, the computation of all models may become impossible. Even worse, it is undecidable in general whether a set of first-order formulas has a model at all. However, Reiter and Mackworth show that as a consequence of the assumptions made in their logical framework, it is possible to enumerate all models. In fact, under the additional closed-world assumption, finite extensions of all predicates can be used in the models, and therefore quantified formulas can be replaced with quantifier-free formulas. Consequently, first-order formulas can be reduced to propositional formulas, for which the computation of all models is possible [GN87]. Reiter and Mackworth formulate the problem of determining all models of the resulting propositional formulas as a *constraint satisfaction problem* (CSP). Although, in general, CSPs of this kind are NP-hard, and thus computationally intractable, several efficient approximation algorithms exist, which have also been used in the Mapsee system [MMH88].

Reiter and Mackworth also show that for the computation of the models using CSP algorithms, only scene axioms are relevant and all other axioms can be ignored. This gives rise to the question whether the distinction between image- and scene-domain knowledge is necessary. This distinction makes the formal specification more involved, but at the same time, allows for a separate representation of general knowledge about the image and scene domains. However, in the first-order logical framework it is not possible to check for the consistency of general knowledge bases, for which no domain-closure axioms can be specified. Furthermore, the logical framework presumes the unambiguous acquisition of image objects, scene objects and their relations, as well as the depiction relations such that unique specifications can be obtained. These assumptions are obviously too strict for general purpose scene interpretation and largely neglect issues such as noise and incompleteness (see also the discussion in [RM90]). Therefore, in [Poo93] Poole, the exploitation of probabilistic knowledge is studied using the Mapsee scenario.

Schröder [Sch98] criticizes that representing interpretation results in terms of logical models (as done in the Mapsee approach) yield interpretations that might be too specific, which, in turn, might cause an over-interpretation of observations. He suggests the notion of a partial model [Sch98], a relational structure detailed enough to represent the commonalities between all models.

1.2.2 Scene Interpretation Based on Abduction

Inspired by the work of Reiter and Mackworth, Matsuyama and Hwang present a vision system called SIGMA, in which they apply logic to scene interpretation [MH90]. In contrast to Reiter and Mackworth, they do not assume the availability of an a priori image segmentation, and do not make domain-closure and unique-name assumptions for the image domain. Constant symbols representing image-domain objects are not available in the beginning, but have to be created through an expectation-driven segmentation approach, which is part of the interpretation process. Consequently, also constant symbols representing scene objects are not available in the beginning of the interpretation process and have to be computed through hypotheses. This is why Matsuyama and Hwang call their approach constructive, and we will argue that nowadays it would have been called *abductive*.

Matsuyama and Hwang use aerial images of suburban areas that typically show houses and roads. First-order logic axioms are used to represent general knowledge about the application domain. For example, the fact that every house is related to exactly one street is represented as follows (for the sake of the example the relation is called *rel*)

$$\forall x : house(x) \Rightarrow (\exists y :\ road(y) \land rel(x,y) \land \forall z : (road(z) \land rel(x,z)) \Rightarrow z = y)$$

This formula can be transformed into *clausal normal form* (with an implicit conjunction operator between the formulas on separate lines).

$$\begin{array}{l}\neg house(x) \lor road(f(x))\\ \neg house(x) \lor rel(x, f(x))\\ \neg house(x) \lor \neg road(z) \lor \neg rel(x,z) \lor z = f(x)\end{array}$$

Existential quantification is replaced with terms using so-called *Skolem functions*. A Skolem term replaces an existentially quantified variable and denotes a certain domain object, depending on the universally quantified variable in whose scope the replaced existentially quantified variable is located. As an example, assume an aerial image depicting a house. The house is represented by the constant h_1. Given the above-mentioned axioms representing the general knowledge about the domain and information about the existence of a house in the scene, namely $house(h_1)$, the following information is entailed:

$$\begin{array}{l}road(f(h_1))\\ rel(h_1, f(h_1))\\ \neg road(z) \lor \neg rel(h_1, z) \lor z = f(h_1)\end{array}$$

Here, the new domain object $f(h_1)$ denoted using the Skolem term f is called an *expected object*, in this example a road, and has to be identified in the image.

In the SIGMA system, different classes of scene objects and spatial relations are defined through necessary conditions.

$$\begin{array}{l}\forall x : road(x) \Rightarrow greater(width(x), 5) \land less(width(x), 100) \land ribbon(shape(x))\\ \forall x, y : rel(x,y) \Rightarrow parallel(axis(x), axis(y)) \land distance(center(x), center(y), 50)\end{array}$$

Object attributes such as *width*, *shape*, *axis* or *center* are modeled through functions, predicates regarding spatial attributes such as *greater*, *less*, *ribbon*, *parallel* or distance are modeled as constraints. These axioms define conditions that must hold for objects of the scene domain, and thus can eliminate certain models.

Assume that our sample image depicts also a road. Then, the road is represented in the scene domain as well, e.g. by the constant r_1. After adding a new axiom to represent this information, namely $road(r_1)$, the following information is entailed:

$$\neg rel(h_1, r_1) \lor r_1 = f(h_1)$$

In the SIGMA system, spatial relations of the image domain are not mapped to relations whose domain and range are the scene domain. In addition, for spatial relations of the scene domain such as *rel* only necessary conditions are defined but not sufficient ones. Therefore, it cannot be proved logically, whether $rel(h_1, r_1)$ holds or not. To solve this problem, a special equality predicate is used in SIGMA, which reflects two important assumptions about equality of objects: i) Two scene objects are considered to be identical, if they are of the same type, e.g. road, and have the same shape and position, i.e. occupy the same space. ii) If an existing object in the scene domain fulfills all conditions that an expected object has to fulfill, both objects are considered to be identical.

In our example, if r_1 fulfills all conditions that have to be fulfilled by the expected object $f(h_1)$ then as a result of the equality assumption, the *hypothesis* $r_1 = f(h_1)$ is generated, and later $rel(h_1, r_1)$ is derived. In case no suitable scene object that is identical to the expected object $f(h_1)$ exists, the conditions of the expected object $f(h_1)$ are used for an expectation-driven image analysis process to identify an object in the image. In case an object is identified, a new constant symbol is introduced in the image domain, e.g. r_2, and the hypothesis $road(r_2)$ is created. Afterwards, the hypothesis $r_2 = f(h_1)$ is generated and $rel(h_1, r_2)$ is derived.

In order to guarantee termination, expected objects are not allowed to trigger the derivation of new expected objects, e.g. $g(f(r_1))$. In other words, expectations are not used to derive further expectations. Expectation generation is done solely through the exploitation of constant symbols, which can only be introduced by an expectation-driven image analysis process.

The hypothesis generation process in SIGMA computes so-called *interpretation networks*, i.e., networks consisting of mutually related object instances. Multiple interpretation networks can possibly be constructed for an image. In an interpretation network, multiple objects instances may be located in the same place in the scene. Such instances are called *conflicting instances*, and a so-called *in-conflict-with* relation is established between them. It should be noted that the SIGMA system applies no heuristics to select among the possible sets of networks but delivers the first computed set of networks as the result.

In [MH90], Matsuyama and Hwang not only present the general approach followed in the SIGMA system, but also discuss the computation of scene interpretations. According to the authors the goal of scene interpretation is to provide for an explanation of the observations, i.e. of the images, through the exploitation of axiomatize general knowledge about the world. They observe that the computation of scene interpretation cannot be achieved through deductive reasoning only: $axioms \not\models observations$.

The axioms representing general knowledge in terms of universally quantified formulas do not entail concrete observations (facts). Instead of a deductive reasoning approach, Matsuyama and Hwang follow the *hypothetical reasoning* approach of Poole et al. [PGA87, Poo89] where the task is to compute a set of logical hypotheses such that following conditions are fulfilled:

i) $Axioms \cup Logical\_Hypotheses \models Observations$
ii) $\mathrm{SAT}(Axioms \cup Logical\_Hypotheses)$

Poole's work is the first in which the space of abducibles is declaratively specified. He uses Horn rules and a set of so-called assumables (aka abducibles) in order to specify which predicates are assumed to be true in a backward-chaining inference process over the Horn rules. The set of these hypotheses are returned as part of the result of the reasoning process (see also [Poo93]). This form of reasoning has been introduced by Peirce [Pei78] under the name *abduction* in the late 19th century. Contrary to deduction where we can reason from causes to effects, in abduction we can reason 'backwards', i.e, from effects (observations) to causes (explanations). Abduction is also often defined as a reasoning process from evidence to explanation, which is a type of reasoning required in several situations where the available information is incomplete [Ali06]. Abduction has been widely used to formalize explanation-based reasoning and plays an important role in intelligent problem solving tasks such as medical diagnosis [PGA87] and plan recognition [CG91]. Formalizing the interpretation of camera data in a robotics scenario, Shanahan has also argued for an explanation-based (abductive) approach to scene interpretation [Sha05]. As described in [Sha05], logic is used for analyzing the behavior of specific procedural programs developed for scene interpretation.

Despite the fact that logic is a useful tool for analyzing (and describing) the behavior of computational systems, and despite the fact that the retrospective use of logic has its merits, nowadays logical reasoning systems have reached a state of maturity such that declarative reasoning services can be used to directly solve the interpretation problems in an abductive way. As has been said before, [Poo93] uses Horn clauses for generating scene interpretations (in an abductive way) and exploits Bayesian networks for ranking alternatives. Recent developments of this significant theory, for which even a practical reasoner implementation exists, can be found in [Poo08] and [PM10].

Another logical approach in which scene interpretation is realized by a practical reasoning engine for ontologies (which are, in some sense, more expressive than Horn clauses) is described in [EKM+07, CEF+49]. This approach has been extended in [GMN$^+$10] in terms of control strategies and w.r.t. ranking explanation probabilities using Markov logic networks. [GMN$^+$10] is the first approach in which the abduction process is systematically controlled by generating an explanation only if the agent can prove that the probability that the observations are true is substantially increased. This solves the termination problem in explanation generation inherent in early approaches such as, e.g., the one of Matsuyama and Huang.

Besides abduction, in [EKM$^+$07] also deduction plays an important role. Something is abduced only if it cannot be proven to hold. We therefore analyze related work on scene interpretation based on deduction. The main question is whether the input (stemming from low-level scene analysis processes) can be made specific enough such that useful conclusions can be computed using deduction principles. Interestingly, somewhat contrary to common expectations, the main message here is, logical deduction is indeed able to compute important results w.r.t. scene interpretation based on sensible expectations w.r.t. analysis results.

1.2.3 Scene Interpretation Based on Deduction

First Approaches to Logic-based Interpretation using Deduction

The VEIL project (Vision Environment Integrating Loom) [RPM$^+$98, RMS97] is a research project that aims to improve computer vision programs by applying formal knowledge representation and reasoning technology. To this end a layered architecture integrating vision processing and knowledge representation has been proposed. In this architecture a computer vision program operates at the pixel level using specialized data structures to deal with low-level processing, whereas the knowledge representation system Loom uses symbolic structures to represent and reason higher-level knowledge.

One of the major goals of VEIL is to enable the construction of explicit declarative vision models. This is achieved by exploiting the knowledge representation and reasoning facilities provided by the Loom system [MB87, Bri93]. The Loom system provides support for an expressive knowledge representation language in the KL-ONE family and reasoning tasks. It supports not only deductive reasoning but provides also facilities to apply production rules. The declarative specification of knowledge offers various benefits: i) It is easier to maintain than a procedurally specified program. ii) It enables the application of automatic validation and verification techniques. iii) Data is represented in a high-level specification instead of application-specific data structures, and thus can easily be shared or reused by other applications.

Similar to the Mapsee and to SIGMA systems, also in the VEIL project domain knowledge is represented in two different models. The *site model* is a geometric model of concrete objects found in a particular image such as runways, markings, buildings and vehicles. The so-called *domain model* contains not only concrete objects such as roads, buildings, and vehicles but also abstract objects such as convoys (groups of vehicles) and field training exercise events.

The VEIL application scenario is the detection and analysis of aerial photographs of airports. Airports are modeled as collections of runways, which are long thin ribbons with markings (smaller ribbons) in certain locations. Aerial images are analyzed by the computer vision system through standard analysis techniques such as the Canny edge detector [Can86] to produce hypotheses. A sequence of filtering and grouping operations are then applied to reduce the number of hypotheses. In the next step, hypotheses are verified using the information in Loom's site model. For example, the site model describes markings in terms of their sizes, relative positions and position on the runway. The domain knowledge represented using Loom is used to constrain the set of possible hypotheses. For example, descriptions of the size and location of markings are used to rule out some hypotheses generated by the computer vision system. The generation of hypotheses, however, is not declaratively modeled. Logic-based deduction (consistency checking) is used to narrow down the space of possible hypotheses.

The work on VEIL shows that declarative representations and deduction as an inference service are useful for scene understanding, although the construction of the space of hypotheses for each scene is not done in terms of logical reasoning in VEIL but using a procedural program. In the VEIL project, deductive reasoning is employed to classify an instance as belonging to a concept. For example, assume that a group of pixels in an image is identified as a vehicle instance v_1 and added to the knowledge base. Further analysis of the same group of pixels might unveil that v_1 has tracks. Adding this information to the knowledge base, Loom classifies v_1 as a *tracked-vehicle* instance, where the concept tracked-vehicle is defined as a subconcept of the concept vehicle. This is possible, because the concept tracked-vehicle is defined with necessary and sufficient conditions, which are all fulfilled by v_1. Note that instance classification has been used even before VEIL in the context of detecting visual constellations in diagrammatic languages (cf. [Haa95, Haa96]).

Ontology-based Interpretation

The exploitation of the ideas behind VEIL in the much more formal context of ontologies has been investigated by Hummel in [HTL07, Hum10]. In her work, Hummel describes a realistic scenario for logic-based traffic intersection interpretation. Based on a crossing model using carefully selected primitives, ambiguity is reduced by "integrating" cues in a logical framework. It is interesting to see how underspecified information derived by low-level analysis processes can be enriched using logical reasoning. In contrast to VEIL, which is based on incomplete reasoning, the work of Hummel uses a sound and complete reasoner and an expressive description language. Hummel found that soundness and completeness are mandatory in order to effectively reduce ambiguity such that (indefinite) cues from analysis processes are condensed to obtain useful interpretation results by deductive interpretation processes.

The overall goal of the system defined by Hummel is to facilitate autonomous car driving through the interpretation of road intersections. To this end, the system is provided as input with sensor data from a camera and a global positioning system (GPS) mounted on a vehicle, as well as with data from a digital map. For each road intersection the system is then requested to answer questions such as ‘Which driving directions are allowed on each lane?’, ’Which of the map’s lanes is equivalent to the vehicle’s ego lane?’, etc. Answering such questions requires reasoning since general regulations of roads and intersections as well as partial and non-complementary information from various sensors about the current situation of the car have to be considered together.

In her work, Hummel investigates appropriate ways for representing relevant scene information in description logics (DLs). Being a decidable subset of first-order logic, DLs are are family of logical representation languages for which highly optimized reasoning systems exist. Terminological knowledge is formalized in terms of terminology (concepts and relations) in a so-called Tbox. Assertional knowledge about particular objects is described in a so-called Abox. For an introduction to DLs see [BCM$^+$03].

For typical classes of scene information she proposes generic DL representations, which she refers to as design patterns. In particular, she presents design patterns for representing sensor data and qualitative scene geometry models in DLs. In the context of road intersection interpretation, different sensor setups are investigated as well. If a still image from a single sensor is interpreted, the unique-name assumption (UNA) should be imposed such that two individuals in the Abox are always interpreted (in the sense of first-order logic) as different objects. However if data is acquired by multiple, non-complementary sensors, objects are detected multiple times, and hence the UNA need not hold. For the multiple sensor setup, Hummel requires the UNA to hold within data acquired by a single sensor only, which she calls the local UNA. She reports the local UNA to have been implemented as a procedural extension that enhances a knowledge base through the application of rules in a forward-chaining way.

Furthermore, Hummel investigates scene interpretation tasks with respect to their solvability through standard deductive DL inference services. These tasks are

1. Object detection, i.e., the discovery of new scene objects
2. Object classification, i.e., the assignment of labels to a detected object
3. Link prediction, i.e., predicting the existence and types of relationships between objects
4. Data association, i.e., the identification of a set of measurements as referring to the same object.

For her experiments Hummel develops a sophisticated Tbox for representing a road network ontology (RONNY), in which the qualitative geometry and building regulations of roads and intersections are specified. Building on these grounds, she describes a case study where the logic-enhanced system solves interpretation tasks using RONNY and sensor data from a stereo vision sensor, a global positioning system, and a digital map. The performance of the system in solving object detection, object classification and data association tasks has been evaluated on a sample set of 23 diverse and complex intersections from urban and non-urban roads in Germany.

She shows that in order solve the object classification task with standard DL inference services, the maximum number of individuals in a scene have to be added a priori to the Abox, which describes the scene. A corresponding design pattern has been proposed in [Hum10]. In fact, if this design pattern is applied, the task of object detection can be reduced to the task of object classification, which can be solved using the so-called Abox realization DL inference service. In a nutshell, Abox realization is a deductive DL inference service that computes for all individuals in an Abox A their most-specific concept names w.r.t. a Tbox T. This way, in a sense, objects are "classified", and the classification determines in terms of symbols (names) what the systems knows about domain objects (see the previous subsection on VEIL).

In contrast to object detection and object classification, Hummel identified that the task of link prediction and data association cannot elegantly be solved using DLs.

In [Hum10], it is shown that the system built through the integration of a deductive DL reasoner and a computer vision system can be used to significantly improve recognition rates of the computer vision system.

1.3 Logic-based Text Interpretation

In a similar way as for scene interpretation, logic-based approaches have been used for text interpretation. In particular, the work of Hobbs et al. in [HSME88, HSM93] has been influential in conceptualizing text interpretation as a problem that requires abduction in order to be solved. They developed a linguistic and knowledge-intensive framework to solve the problem of text interpretation starting from the derivation of the so-called logical form of a sentence, a first-order representation capturing its logical structure, together with the constraints that predicates impose on their arguments. The central idea of Hobbs et al. is to show that logical forms of (parts of) sentences can be established as consequences from background knowledge and additional assumptions (formulae to be added). The added formulae provide for a deeper interpretation.

As an example, consider the following sentence, on which a sequence of interpretation steps are applied.

(1) Disengaged compressor after lube-oil alarm.

The *reference resolution* step analyzes the words “compressor” and “alarm” and identifies them as so-called references. To establish the reference of compressor, the following logical form is generated for this part of the sentence

(2) $\exists x : compressor(x)$

Given a background knowledge base containing,

$starting\_air\_compressor(c_1)$
$\forall x : starting\_air\_compressor(x) \Rightarrow compressor(x)$

i.e., an instance of a “starting air compressor”, namely c_1, and the definition of starting air compressor as a specific type of compressor, then the logical form (2) extracted from the sentence (1) can be resolved to the instance c_1, i.e.,

$compressor(c_1)$

is derived, and in this sense, the entailment of expression (2) is proved. In this case no additional assumptions are required.

When a reference formula cannot be proved to be entailed (w.r.t. the background knowledge), then it is assumed to be true. Here, we find the principle of abduction be applied. For example, “Lube-oil alarm” is a *compound nominal*, thus composed of two entities which are implicitly related to each other. The problem of determining the implicit relation between the two is called compound nominal resolution. To interpret “lube-oil alarm”, a logical form is first extracted, namely

$\exists y, z, nn : lube\_oil(ys) \wedge alarm(z) \wedge nn(y, z)$,

and, due to the principle explained above, w.r.t. the background knowledge, it should be possible to find one entity for lube-oil and another for alarm, and there must be some implicit relation (called *nn*) between them. If the entailment of the above formulae cannot be shown, assumptions are necessary (possible with Skolem terms, see above). Note, however, that assumptions need not be “least-specific”. For instance, if the background knowledge contains information about the most common possible relations for an implicit relation, e.g. to denote part-whole relations,

(3) $\forall x, y : part(x, y) \Rightarrow nn(x, y)$

or complex relations that can be explained as a *for* relation,

(4) $\forall x, y : for(x, y) \Rightarrow nn(x, y)$

an assumption using *part* or *for* can in principle be made rather than use the more “abstract” relation directly.

As can be observed, there might exist more than one possibility to make assumptions. To choose between possible candidates, [HSM93] defines a preference strategy to support this decision problem, called weighted abduction which will be explained below. We first continue with the example.

Deciding whether “after lube-oil alarm” modifies the compressor or the disengaging event is the problem of *syntactic ambiguity resolution.* To solve this problem, Hobbs et al. propose the transformation of the problem to a constrained coreference problem, where the first argument of the predicate is considered as existentially quantified. In this sense, the extracted logical expression is:

$$(5)\ \exists e, c, y, a : \mathit{after}(y, a) \land y \in \{c, e\}$$

where the existentially quantified variable y should be resolved to the compressor c or the disengaging event e. This problem is often solved as a by-product of metonymy resolution. *metonymy resolution* which involves the “coercion” of words such that the constraints that predicates impose on their arguments are fulfilled.

For example, in the above sentence (1), the predicate *after* requires events as arguments:

$$(6)\ \forall e_1, e_2 : \mathit{after}(e_1, e_2) \Rightarrow \mathit{event}(e_1) \land \mathit{event}(e_2).$$

Therefore, it is necessary to coerce the logical form in (5) such that the requirements of the predicate *after* are fulfilled. For this purpose, coercion variables satisfying the constraints are introduced:

$$\begin{aligned}(7)\ \exists k_1, k_2, rel_1, rel_2, y, a : {} & \mathit{after}(k_1, k_2) \land \mathit{event}(k_1) \land rel_1(k_1, y) \land \\ & \mathit{event}(k_2) \land rel_2(k_2, a)\end{aligned}$$

in this case k_1 and k_2 are the coercion variables related to *after* instead of y and a as it was before. Also coercion relations (rel_1, rel_2) are introduced. As can be seen from the example, coercion variables and relations are implicit information and are also generic, which suggests that any relation can hold between the implicit and the explicit arguments. If there are axioms in the background knowledge base, expressing the kind of “coercions” that are possible:

$$\forall x, y : \mathit{part}(x, y) \Rightarrow rel(x, y)$$
$$\forall x, e : \mathit{function}(e, x) \Rightarrow rel(e, x)$$

then, metonymy resolution is solved by abduction.

The next phase aims at computing the cost of the resulting interpretation. It is anticipated that during the process of proving the entailment of a logical form (see above) different proofs can be found. In order to find the “less-expensive” proof Hobbs et al. developed a method called weighted abduction, which is characterized by the following three features: First, goal expressions should be assumable at varying costs, second it should be possible to make assumptions at various levels of specificity, and third, natural language redundancy should be exploited to yield more economic proofs. In this method, each atom of the resulting ungrounded logical form is weighted with a cost. For instance, in the formula

$$\exists e, x, c, k_1, k_2, y, a, o : Past(e)^{\$3} \wedge disengage'(e, x, c)^{\$3} \wedge compressor(c)^{\$5} \wedge after(k_1, k_2)^{\$3} \wedge event(k_1)^{\$10} \wedge rel(k_1, y)^{\$20} \wedge y \in \{c, e\} \wedge event(k_2)^{\$10} \wedge rel(k_2, a)^{\$20} \wedge alarm(a)^{\$5} \wedge nn(o, a)^{\$20} \wedge lube\_oil(o)^{\$5}$$

costs are indicated as superscripts with $ signs. Costs indicate different weights. An explanation is preferred if the costs of the things to assume are minimal.

The costs are given according to linguistic characteristics of the sentence, thus if the same sentences is expressed in a different way, the cost might vary accordingly. They have analyzed how likely it is that a linguistic expression conveys new information, and therefore failing to prove the entailment of the construct is not so costly, contrary to other expressions in which no new information is conveyed, and therefore it should be possible to prove the corresponding entailment. For example, the main verb is more likely to convey new information than a definite noun phrase which is generally used referentially. Failing to prove a definite noun phrase is therefore expensive. For a more detailed description of this linguistic characteristics, refer to [HSM93].

Besides these weights, there are other factors used to determine the appropriateness of an interpretation, namely simplicity and consilience. A simple interpretation would be one that exploits redundancy in the discourse, such that the number of assumptions can be reduced, for example by assuming that two atoms are identical due to semantic knowledge. Consilience refers to the relation between the number of atoms that have been proved exploiting redundancy and the total number of atoms to prove. The highest the number of atoms that have been proved with the less number of assumptions, the more the explanation is consilient.

In their approach, less-specific explanations are preferred, due to the fact that the more specific the assumptions are, the more information can be obtained but it is also more likely that they are not correct. This is the so called informativeness-correctness trade-off. However, if there is evidence in the background knowledge that allows for a more specific assumption, then the more specific proof is considered.

As we have argued, the work of Hobbs et al. show us that logic-based interpretation can account for a large number of effects that naturally occur in text interpretation. We are now ready to study another modality, namely the video modality.

1.4 Logic-Based Video Interpretation

For video interpretation, various ontologies have been used. Whereas in some approaches time points are used (with time points being specified by quantitative numerical values), other approaches use time intervals and qualitative relations between them. What distinguishes the approaches is the level of declarativeness of how events to be recognized are specified.

1.4.1 Early Approaches

The beginnings of symbolic video interpretation can be dated back to the seminal publication of Tsotsos et al. [TMCZ80] describing the ALVEN system for automatic heart disease detection. The basic idea of ALVEN is to use a frame-based representation in which each frame can be associated with spatio-temporal constraints describing instantiation restrictions. Spatio-temporal motion phenomena such as heart contractions are described in terms of area changes (the initial area is larger than the resulting area). The change can, for instance, be further characterized using a speed specification, which can be further constrained using additional predicates describing necessary conditions (e.g., the area change must not be too large or too small). A small set of primitive movement descriptors, such as time interval, location change, length change, area change, shape change, and so on are used to describe all higher-level motion concepts. Event frames can be linked to one another using so-called similarity links. Different techniques for event recognition and hypothesis ranking are explored. The description language used in ALVEN is inspired by natural language descriptions for motion events investigated in [Bad75].

Although ALVEN uses a procedural description for the event recognition process, and does not model event recognition as a logical reasoning problem (besides inheritance reasoning), it was one of the first systems to use explicit symbolic representations. ALVEN has influenced the work of Neumann et al. who were among the first to use a logic-based approach for recognizing events in street scenes.

1.4.2 Quantitative Approaches for Event Definitions

The goal of Neumann and Novak [NN83, Neu85, NN86] was to support query answering and the generation of natural language descriptions for street scene events (the system was called NAOS: NAtural language description of Object movements in Street scenes). The basis for the NAOS system is a so-called geometric scene description (GSD): Per timepoint the description consists of detected objects including their types and their positions.

Given a GSD determined by low-level video analysis processes, basic motion event descriptions of single objects are generated. Basic motion events such as move, accelerate, approach, etc. are associated with two timepoints (start point and end point) in such a way that the resulting interval is maximal w.r.t. a sequence of GSD snapshots. Given a set of assertions for basic motion events, high-level motion events are instantiated based on a set of declarative event models. The following example demonstrates the main ideas behind NAOS.[4]

[4] The original syntax used in the NAOS system slightly deviates from the example presented here. We describe the syntax used in a reimplementation of the NAOS event recognition engine, which is based on the work described in [MN08].

```
(define-event-class ((overtake ?obj1 ?obj2) *t1 *t2)
   (?obj1 vehicle)
   (?obj2 vehicle)
   ((move ?obj1) *t1 *t2)
   ((move ?obj2) *t1 *t2)
   ((approach ?obj1 ?obj2) *t1 *t3)
   ((behind ?obj1 ?obj2) *t1 *t3)
   ((beside ?obj1 ?obj2) *t3 *t4)
   ((in-front-of ?obj1 ?obj2) *t4 *t2)
   ((recede ?obj1 ?obj2) *t4 *t2))
```

Events are specified as Horn rules with object variables (indicated with ?) and time variables (prefixed with *). The first two conditions impose non-temporal static restrictions on the types of ?obj1 and ?obj2. The temporal relation between subevents are indicated using corresponding time variables. See [MN08] for a detailed definition of the semantics of event classes in terms of logical rules.

Implicit constraints are established between temporal variables. We give the semantics of the above definition in CLP($\mathfrak{R}$) [JMSY92], where *holds*(*Atom*) means that *Atom* can be proven using an external prover, in this case a description logic reasoner.

```
overtake(Obj1, Obj2, T1, T2) :- T1 < T2,
   holds(vehicle(Obj1)),
   holds(vehicle(Obj2)),
   move(Obj1, T1, T2), T1 < T2,
   move(Obj2, T1, T2), T1 < T2,
   approach(Obj1, Obj2, T1, T3), T1 < T3,
   behind((Obj1, Obj2, T1, T3), T1 < T3,
   beside((Obj1, Obj2, T3, T4), T3 < T4,
   in_front_of((Obj1, Obj2, T4, T2), T4 < T2,
   recede(Obj1, Obj2, T4, T2), T4 < T2.
```

An example for a set of basic motion events derived from a GSD is given below (we use constants vw1 and vw2).

```
(define-assertion ((move vw1) 7 80))
(define-assertion ((move vw2) 3 70))
(define-assertion ((approach vw1 vw2) 10 30))
(define-assertion ((behind vw1 vw2) 10 30))
(define-assertion ((beside vw1 vw2) 30 40))
(define-assertion ((in-front-of vw1 vw2) 40 80))
(define-assertion ((recede vw1 vw2) 40 60))
```

With (define-assertion ((R X) T1 R2)) a corresponding CLP($\mathfrak{R}$) fact R(X, T1, T2). is denoted (analogously for (R X Y)).

An example query for our scenario is given as follows (with query results printed below in terms of variable bindings).

```
(?- ((overtake ?obj1 ?obj2) *t1 *t2))
-> OBJ1 = VW1 OBJ2 = VW2 T1 = [10, 29] T2 = [41, 60]
```

The substitutions for object and time variables indicate recognized events. Time intervals indicate the minimum and maximum values for which the event can be instantiated given the assertions for basic motion events specified above. It is also possible to query for events involving specific objects (e.g., `vw2`).

```
(?- ((overtake ?obj1 vw2) *t1 *t2))
-> OBJ1 = VW1 T1 = [10, 29] T2 = [41, 60]
```

Note that in contrast to CLP($\mathfrak{R}$), in NAOS there are actual solutions being generated, and not only consistency checks performed for time variables (or real variables). Given a query (goal specification), NAOS applies some form of backward chaining of event class rules to determine bindings for variables. Backward chaining involves constraint propagation for time variables [Neu85].

In NAOS it is also possible to find instantiations of all declared event models. The rules are applied in a forward chaining way if there is no specific goal.

```
(?-)
Rule OVERTAKE indicates ((OVERTAKE VW1 VW2) 10 60).
```

Based on the bindings found for events, it is possible to explicitly add event assertions to the knowledge base (e.g., `define-assertion ((overtake vw1 vw2) 10 60)`. These assertions can then be used to detect even higher-level events.

As can be seen from the example, the original NAOS system can be used for an a-posterior analysis of a given set of event assertions. In principle, the approach can be extended to support incremental event recognition (see [KNS94] for an early approach based on quantitative data) such that one can also query for events which might be possible at a certain timepoint.

It should also be emphasized that in general there might be multiple possibilities for instantiating events. Thus, a combinatorial space for navigating through logic-based scene models is defined (see [NW03] for details). Scene models define classes for high-level events using a first-order language. The construction process for valid interpretation hypotheses described in [NW03] is extra-logical, however.

Horn clauses are not the only logical representation language that has been used in the literature to specify events. In an attempt to formulate scene understanding and event recognition as a (sequence of) logical decision problem(s), the approach described in [NM06] uses ontologies (aka description logic knowledge bases) as the underlying formalism. In particular, high-level event descriptions are generated by employing ontology-based *query answering*. The approach in [NM06] does not specify, however, from which knowledge sources the queries are taken. Along the same lines, a more methodological approach is presented by the same authors in [MN08]. In this work, event-query generation is formalized as a form of abductive reasoning, and the space of abducibles is defined by rules.

As we have seen, in the NAOS approach, event recognition is based on quantitative information on timepoints, and (simple) constraints over the reals ensure the semantics of time to be represented in NAOS. In addition, event assertion with maximal time intervals must be made available to the NAOS formalism. All assertions are maintained in a large knowledge base.

1.4.3 Qualitative Approaches for Event Definitions

Another idea to represent events is to subdivide facts into temporally ordered partitions and use qualitative relations between the partitions. This has been explored in the VEIL system (see above) in order to detect event sequences that span multiple images. The goal of this scenario is to process a sequence of images and detect events such as field training exercises. Forty images of a hypothetical armored brigade garrison and exercise area that share a common site model have been used in the experiments reported in [RPM^{+}98].

In the VEIL context, an event is a sequence of scenes that satisfy certain criteria. A scene is represented as a set of object descriptions (called a world), which can be associated with a timestamp. Some of the criteria such as the temporal order apply across different scenes, whereas other criteria apply only within a single scene.

A field training exercise is a sequence of scenes showing an armored unit in a garrison, then moving in convoy, then deployed in a training area and finally in a convoy again. In order to extract the scenes that meet the criteria of a field training exercise event, the following query is used:

```
(retrieve (?Y ?S1 ?S2 ?S3 ?S4)
      (and (within-world ?S1 (in-garrison ?Y))
           (within-world ?S2 (convoy ?Y))
           (within-world ?S3 (deployed-unit ?Y))
           (within-world ?S4 (convoy ?Y))
           (before+ ?S1 ?S2)(before+ ?S2 ?S3)(before+ ?S3 ?S4)))
```

Query terms, e.g. *in-garrison* and *deployed-unit*, are defined in the domain model. The result of the query is a set of tuples. Each tuple is a field training exercise event since it satisfies all conditions defined in the query.

It should be pointed out that qualitative relations between states (worlds) are used in the query language. In this context, there are means for adding specification of events to the Tbox (see, e.g., [ALT07]).

In all approaches to image sequence understanding, be they quantitative or qualitative, it is important to understand what is made explicit and what is added by logical reasoning. The corresponding “design space” is investigated in detail in [WLM09]. The main insight is that, given the features of contemporary reasoning systems, event recognition can be formalized as query answering in the expressive description logic Abox query language nRQL (for an introduction to nRQL see [WM05]). Building on this query language, [BBB$^+$09] formalize event recognition (actually, in [BBB$^+$09] event recognition is called situation recognition) by transforming specifications of linear temporal logic into nRQL queries.

1.5 Summary

We have sketched major logic-based representation languages that formalize interpretation using logical decision processes. The most important insights gained from these works are:

- Existing computer vision systems are well-equipped to process pixel-level data, whereas formal knowledge representation and reasoning systems are more appropriate to process symbolic structures. Therefore it is reasonable to distinguish between surface-level and deep-level information when building a software system for scene interpretation.
- Even though a scalable system for declarative scene interpretation could not be built yet, promising results have been achieved. Various benefits such a system would offer motivate us to develop future logic-based approaches for multimedia interpretation.
- It is hardly possible to compute interpretations of an image through deductive reasoning only. The generation of hypothesis in an abductive way is crucial for scene interpretation, and provides for an appropriate formalization of the generative nature of the interpretation process. Representing interpretation results in terms of logical models (see the Mapsee approach) seems to be too specific, and the specificity of models provides for an over-interpretation of observations.

The goal of scene interpretation is to provide for explanations of the observations made through the analysis of an image. The explanations have to be hypothesized since, in general, observations are not entailed by available background knowledge. In fact, if the available background knowledge would contain explanations of observations, the computation of an scene interpretation would be unnecessary since the scene interpretation would already be part of the background knowledge. Therefore the observations can logically follow from the background knowledge only if appropriate explanations are hypothesized and added to the background knowledge before. Different approaches exist in the literature for specifying the “space of abducibles”.

References

[AHB+93] D. E. Appelt, J. R. Hobbs, J. Bear, D. J. Israel, and M. Tyson. FASTUS: A Finite-state Processor for Information Extraction from Real-world Text. In *Proceedings of IJCAI*, pages 1172–1178, 1993.

[Ali06] A. Aliseda. *Abductive Reasoning: Logical Investigations into Discovery and Explanation*, volume 330 of *Synthese Library*. Springer, 2006.

[ALT07] Alessandro Artale, Carsten Lutz, and David Toman. A description logic of change. In Manuela Veloso, editor, *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 218–223. AAAI Press, 2007.

[Bad75] N. Badler. Temporal scene analysis: Conceptual descriptions of object movements, report tr-80. Technical report, Dept. of CS, University of Toronto, 1975.

[BBB+09] Franz Baader, Andreas Bauer, Peter Baumgartner, Anne Cregan, Alfredo Gabaldon, Krystian Ji, Kevin Lee, David Rajaratnam, and Rolf Schwitter. A novel architecture for situation awareness systems. In Martin Giese and Arild Waaler, editors, *Proceedings of the 18th International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (Tableaux 2009)*, volume 5607 of *Lecture Notes in Computer Science*, pages 77–92. Springer-Verlag, 2009.

[BCM+03] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press, 2003.

[BCSW07] Holger Bast, Alexandru Chitea, Fabian Suchanek, and Ingmar Weber. Ester: efficient search on text, entities, and relations. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2007)*, pages 671–678, 2007.

[BMP02] S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 24(24), pages 509–522, 2002.

[Bri93] D. Brill. *Loom Reference Manual.* Information Sciences Institute, University of Southern California, 4676 Admiralty Way, Marina del Rey, CA 90292, December 1993.

[Can86] J. F. Canny. A Computational Approach To Edge Detection. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 8(6):679–698, November 1986.

[CEF$^+$49] S. Castano, S. Espinosa, A. Ferrara, V. Karkaletsis, A. Kaya, R. Möller, S. Montanelli, G. Petasis, and M. Wessel. Multimedia Interpretation for Dynamic Ontology Evolution. *Journal of Logic and Computation*, Advance Access published on September 30, 2008. doi:10.1093/logcom/exn049.

[CG91] E. Charniak and R. Goldman. Probabilistic Abduction For Plan Recognition. Technical report, Brown University, Tulane University, 1991.

[CY99] S. Cucerzan and D. Yarowsky. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In *Proceedings of Joint SIGDAT Conf. on Emprical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[EKM$^+$07] S. Espinosa, A. Kaya, S. Melzer, R. Möller, and M. Wessel. Towards a Media Interpretation Framework for the Semantic Web. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, pages 374–380, Washington, DC, USA, November 2007. IEEE Computer Society.

[GMN$^+$10] Oliver Gries, Ralf Möller, Anahita Nafissi, Maurice Rosenfeld, Kamil Sokolski, and Michael Wessel. A probabilistic abduction engine for media interpretation. In J. Alferes, P. Hitzler, and Th. Lukasiewicz, editors, *Proc. International Conference on Web Reasoning and Rule Systems (RR-2010)*, 2010.

[GN87] M. R. Genesereth and N. J. Nilsson. *Logical Foundations of Artificial Intelligence.* Morgan Kaufmann Publ. Inc., Los Altos, CA, 1987.

[Gri03] R. Grishman. Information Extraction. In *Handbook of Computational Linguistics Information Extraction*, 2003.

[Haa95] V. Haarslev. Formal semantics of visual languages using spatial reasoning. In *Proceedings of the 11th IEEE Symposium on Visual Languages, Sept. 5-9, Darmstadt, Germany*, pages 156–163. IEEE Press, 1995.

[Haa96] Volker Haarslev. A fully formalized theory for describing visual notations. In *Proceedings of the AVI'96 post-conference Workshop on Theory of Visual Languages, May 30, 1996, Gubbio, Italy*, 1996.

[HS88] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Proceedings of 4th Alvey Vision Conference*, pages 147–151, 1988.

[HSM93] J. R. Hobbs, M. Stickel, and P. Martin. Interpretation as abduction. *Artificial Intelligence*, 63:69–142, 1993.

[HSME88] J. Hobbs, M. Stickel, P. Martin, and D. Edwards. Interpretation as Abduction. In *26th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, 1988.

[HTL07] B. Hummel, W. Thiemann, and I. Lulcheva. Description logic for vision-based intersection understanding. In *Proc. Cognitive Systems with Interactive Sensors (COGIS), Stanford University, CA*, 2007.

[Hum10] Britta Hummel. *Description Logic for Scene Understanding at the example of Urban Road Intersections.* Südwestdeutscher Verlag für Hochschulschriften, 2010.

[JB08] Y. Jing and S. Baluja. PageRank for Product Image Search. In *Proceedings of 17th International World Wide Web Conference WWW 2008*, April 2008.

[JMSY92] Joxan Jaffar, Spiro Michaylov, Peter J. Stuckey, and Roland H. C. Yap. The CLP(R) language and system. *ACM Transactions on Programming Languages and Systems*, 14(3):339–395, 1992.

[KLSG03] B. Katz, J. Lin, C. Stauffer, and E. Grimson. Answering Questions About Moving Objects in Surveillance Videos. In *Proceedings of AAAI Spring Symposium on New Directions in Question Answering*, March 2003.

[KNS94] S. Kockskämper, B. Neumann, and M. Schick. Extending process monitoring by event recognition. In *Proc. Second International Conference on Intelligent System Engineering ISE-94*, pages 455–460, 1994.

[Low04] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. In *International Journal of Computer Vision*, volume 60(2), pages 91–110, 2004.

[LS03] B. Leibe and B. Schiele. Interleaved Object Categorization and Segmentation. In *Proceedings of British Machine Vision Conference (BMVC'03)*, September 2003.

[MB87] R. M. MacGregor and R. Bates. The Loom Representation Language. Technical Report ISI/RS-87-188, Information Sciences Institute, University of Southern California, 1987.

[MH90] T. Matsuyama and V. S. Hwang. *SIGMA: A Knowledge-Based Aerial Image Understanding System.* Perseus Publishing, 1990.

[MMH88] J. A. Mulder, A. K. Mackworth, and W. S. Havens. Knowledge structuring and constrataint satisfaction: The Mapsee approach. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 10(6):866–879, 1988.

[MN08] R. Möller and B. Neumann. Ontology-based Reasoning Techniques for Multimedia Interpretation and Retrieval. In *Semantic Multimedia and Ontologies : Theory and Applications*. Springer, 2008.

[MS05] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Desciptors. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 27(10), pages 1615–1630, 2005.

[Neu85] Bernd Neumann. Retrieving events from geometrical descriptions of time-varying scenes. In J.W. Schmidt and Costantino Thanos, editors, *Foundations of Knowledge Base Management – Contributions from Logic, Databases, and Artificial Intelligence*, page 443. Springer Verlag, 1985.

[NM06] B. Neumann and R. Möller. On Scene Interpretation with Description Logics. In *Cognitive Vision Systems: Sampling the Spectrum of Approaches*, volume 3948 of *Lecture Notes in Computer Science*, pages 247–278. Springer, 2006.

[NN83] Bernd Neumann and Hans-Joachim Novak. Event models for recognition and natural language description of events in real-world image sequences. In *Proc. International Joint Conference on Artificial Intelligence, IJCAI-83*, pages 724–726, 1983.

[NN86] B. Neumann and H.-J. Novak. NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenenxs. *Informatik Forschung und Entwicklung*, 1:83–92, 1986.

[NW03] 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 212–22. Springer, 2003.

[Pei78] C. S. Peirce. Deduction, Induction and Hypothesis. In *Popular Science Monthly 13*, pages 470–482, 1878.

[PGA87] D. Poole, R. Goebel, and R. 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*, pages 331–352. Springer, 1987.

[PM10] D. Poole and A. Mackworth. *Artificial Intelligence: foundations of computational agents*. Cambridge University Press, New York, NY, 2010.

[Poo89] D. Poole. Explanation and Prediction: An Architecture for Default and Abductive Reasoning. *Computational Intelligence*, 5(2):97–110, 1989.

[Poo93] David Poole. Probabilistic horn abduction and bayesian networks. *Artificial Intelligence*, 64(1):81–129, 1993.

[Poo08] David Poole. *Probabilistic Inductive Logic Programming: Theory and Application*, volume 4911 of *LNAI*, chapter The independent choice logic and beyond. Springer Verlag, 2008.

[RM87] R. Reiter and A. K. Macworth. The Logic of Depiction. Technical Report 87-24, Department of Computer Science, University of British Columbia, Vancouver, Canada, 1987.

[RM90] R. Reiter and A. K. Macworth. A Logical Framework for Depiction and Image Interpretation. *Artificial Intelligence*, 41(125-155), 1989/90.

[RMS97] T. A. Russ, R. M. MacGregor, and B. Salemi. VEIL: Combining Semantic Knowledge with Image Understanding. In O. Firschein and T.M. Strat, editors, *Radius: Image Understanding for Imagery Intelligence*, pages 409–418, San Francisco, CA, 1997. Morgan Kaufmann.

[RPM$^+$98] T. Russ, K. Price, R. M. MacGregor, R. Nevatia, and B. Salemi. VEIL: Research in Knowledge Representation for Computer Vision, Final Report. Technical Report A051143, Information Sciences Institute, University of Southern California, February 1998.

[Sch98] C. Schröder. *Bildinterpretation durch Modellkonstruktion: Eine Theorie zur rechnergestützten Analyse von Bildern*. PhD thesis, University of Hamburg, 1998.

[Sha05] M. P. Shanahan. Perception as Abduction: Turning Sensor Data Into Meaningful Representation. *Cognitive Science*, 1:103–134, 2005.

[SHB07] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. Thomson Learning, April 2007.

[SLL02] S. Se, D. Lowe, and J. J. Little. Global Localization using Distinctive Visual Features. In *Proceedings of International Conference on Intelligent Robots and Systems (IROS2002)*, pages 226–231, Lausanne, Switzerland, November 2002.

[SS08] C. Saathoff and S. Staab. Exploiting spatial context in image region labelling using fuzzy constraint reasoning. In *Ninth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-08)*, pages 16–19, 2008.

[TMCZ80] J.K. Tsotsos, J. Mylopoulos, H.D. Covvey, and S.W. Zucker. A framework for visual motion understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1980.

[VJ01] P. Viola and M. Jones. Robust Real-time Object Detection. In *International Journal of Computer Vision*, 2001.

[WLM09] M. Wessel, M. Luther, and R. Möller. What happened to Bob? Semantic data mining of context histories. In *Proc. of the 2009 International Workshop on Description Logics DL- 2009, 27 to 30 July 2009, Oxford, United Kingdom*, 2009. CEUR Workshop Proceedings (Vol. 477).

[WM05] M. Wessel and R. Möller. A high performance semantic web query answering engine. In I. Horrocks, U. Sattler, and F. Wolter, editors, *Proc. International Workshop on Description Logics*, 2005.