Where do Logic and Computer Vision meet?

Marco Aiello and Arnold Smeulders

Abstract

In this position paper, after giving a short overview of computer vision, we argue for the need of considering applications while devising theories that deal with space. We present the spatial coherence constraint, which is imposed by nature on vision. Finally, we suggest a possible bridge to connect computer vision with spatial reasoning.

Contents

1 Foreword

When I (Marco) read the posting from the University of Amsterdam for a PhD position supervised by Johan van Benthem and Arnold Smeulders I was quite surprised at first. What could bring together two such outstanding, but different researchers?

Surprise led to curiosity; curiosity led to challenge; and all in all I decided to apply for the position. I knew that the groups led by Johan and Arnold were both doing excellent research, so I hoped that working within them would be an additive process on respective successes.

To my further surprise, when I started approaching the problem of spatial reasoning and reasoning with images, I soon realized that, even though Johan and Arnold had very different points of view, they had *real* ideas on the topic. The research position I filled was not only generated by the will of having a joint project, it was really generated by deep thoughts.[1]

In this paper, Arnold and I give an introduction to computer vision, assuming the reader more familiar with logic and formal methods than computer vision itself. We then bring the attention to constraints on perception imposed by nature. Constraints that should be taken into account when considering formal theories of perception. We conclude the paper sketching were logic and computer vision can possibly meet in order to have highly effective vision systems.

We gladly dedicate this paper to Johan on his 50^{th} birthday.

2 Computer Vision

In this section[2] we give some definitions in computer vision, intending to provide a brief introduction to the field and trying to stress the fact that most of the techniques are specialized syntactic manipulators. With this we mean that the state of the art in computer vision is mostly about syntactic properties of images.

- A "raw" **image** is the result of a physical observation of some field, it consists of a matrix of values that captures the field's coherent structure. We use the word raw to stress the fact that the image is the result of a sampling process of a dense physical field, usually, but not necessarily, electromagnetic. More specifically, a digital image is a matrix of values.

  The probing sensor may have any relationship to the object in view. Hence, translation, rotation, scale projection and occlusion are important classes of variance in visual probing.

- **Image Processing** functions map image data fields into another data field. Thus images are the input and the output of image processing (IP

[1] The work presented in this position paper comes from the first year of research carried out within the above mentioned project. The ideas underlying the last section are joint work with Johan van Benthem.

[2] The content of this section is partially based on Section 2.1 of [SKG98].

from now on). Common purposes for which these functions are designed are *restoration:* the process of eliminating, as far as possible, the distortion and noise introduced by the probing sensors; *enhancement:* the highlighting of elements of task-specific interest in an image; *optic flow:* pixel velocity estimation in a sequence of images; *compression:* the reduction of the quantity of data with or without loosing the least relevant details.

- **Segmentation** methods are computational methods that *segment* the image into regions, where each region should be mapped to a physical object or location in the scene. In other words, the set of pixels is partitioned into meaningful entities. Many different computational techniques for segmentation exist, none of which is capable of handling any reasonable set of real world images, but they can behave well in strongly limited domains. As the definition expresses—i.e. establishing a relationship between the computational side and the physical world—defining a general theory of segmentation will be from difficult to impossible. Illumination, shadows, and surface properties are additional classes of variance here.
- A **feature** captures some aspect of an image, a point in the image, a patch in the image, or—after segmentation—an object in the image.
- A **histogram of feature values** of an image is the frequency range of feature values in the image; for example the statistics of all color values of every pixel in the image.

These definitions should make clear what kind of ontologies are used in computer vision; the stress is on the fact that state of the art computer vision deals mostly with syntactic properties of images. **Symbolic model-based image analysis** combines AI type reasoning with vision modules which provide observations. Available systems today are most suited for map-interpretation. **Geometrical model-based image analysis** is about semantics, but limited in scope. This kind of analyzers have models for a very restricted domain, may that be the structure of the human ankle bone or, for example, houses in the district of Bonn.[3]

As the examples of geometrical model-based image analysis show, the semantics involved in computer vision concern mostly elementary properties of the objects depicted in the images, such as the geometrical properties. Borrowing terminology from natural language processing, we could say that we lack a *parser* for images, or more precisely, general parsers. It is hard to find a grammar for images, it is hard to find a suitable intermediate or internal form to represent images. On the other hand, even if we had semantic representations of images we would run into difficulties, this because we still lack "semantic analyzers."

If the above problem has found elegant solutions in the field of artificial languages, such as programming languages, and reasonable solutions are beginning

[3]Many of these analyzers require human interaction to provide successful analysis. In the examples given in the text, a physician with expertise in the reading of an X-ray photograph or a geographer with experience with aerial photographs would interact with the specific computer vision system.

to arise in the more general field of natural language processing, we are still far from achieving similar results in computer vision. It is quite easy to see that solving the "computer vision problem" has the same difficulties and complexity involved in achieving real general AI.

After having presented concepts from two subfields of computer vision, namely Image Processing and Image Analysis, in the next section, we turn to Image Understanding, getting closer to the logical point of view of the subject.

2.1 The Challenge of Real World Applications

In the early days of AI, vision was considered an easy problem (as noted by Marr in [Mar83]) and thus was not attacked properly. Algorithms that were supposed to do "vision" were devised, considering as input simplified representations of the world. It was the time of the block worlds, that indeed were extremely useful as benchmarks, but were absolutely inadequate as tests for realistic applications. The main drawback of the use of "simplified versions of the world," is that it may keep researchers away from considering some of the key aspects.

If one assumes that identifying a cube in an image is only about grouping together 9 segments, one is then maybe inclined to think that vision is a simple task; but he has neglected at least two important steps. On the IP side, extracting the segments, or more properly the edges from an image is, in general, a non trivial task; on the knowledge representation side, one cannot assume that the world is made of homogeneous, simply placed, boxes, but has to take into account that the world is formed of objects which are geometrically complex, with very different surface textures, and that can be viewed from an infinite number of positions.

A mistake that sometimes still shows itself when a new problem is brought to the attention of the communities of computer vision or logic, is the slogan "I can see it, so a computer can see it too" accompanying it. If a human being can discern a spatial situation, it means that there is an observable difference. Identifying this difference and, even more, encoding this difference formally, is in general non trivial.

For example, take the letters that compose this instance[4] of the word "Johan." Identifying the five letters is routine work for any human with basic education, but it took more than 20 years to have this task performed by a computer for printed fonts! Once the Optical Character Recognition (OCR) problem for printed letters found adequate solutions, researchers started to tackle the same issue for handwritten texts. Some researchers that were able to read their own handwriting, following the above mentioned slogan, promised solutions within a couple of years. Time has passed, but up to now OCR for handwriting is possible only adapting the users handwriting to the OCR's requirements.

Our conclusion regarding past experience is that one should avoid devising a theory about vision or spatial reasoning without considering from the very

[4]We mean the word as typed on the paper in front of you, or that you would find on the printed version of the electronic page you are reading right now.

beginning applications to real world problems. Nature imposes constraints on the human perceptual process, and these constraints also drive the way we reason about space.

In the next section we identify the *spatial coherence constraint*, which we believe is what still keeps far apart most of the work in spatial reasoning from computer vision and vice versa. This constraint is evident only when working with real physical quantities, thus justifying our claim that there is a need for application in the spatial reasoning community in order to fill the enormous gap that separates it from computer vision.

3 The Intrinsic Nature of Vision

All perceptual processes involve physical entities to percept and probing sensors. Vision is no exception.

The way one percepts, for instance, a flower depends on which probing sensors are used. If one uses only the right thumb one can discern the texture roughly, and have just a vague idea of its dimension. Using olfaction one can tell if the flower has a good scent or not, and one can possibly discover what kind of flower it is. Using sight, the amount of information that one gathers is much bigger, but always tied to the probing sensor. The eye (and the underlying neural material) is a collection of very sophisticated sensors, the range of discernible differences is very high, but not infinite. One cannot "see" the temperature of the flower, or distinguish differences in color beyond a certain limit.

Perception is constrained as all the acts of sensing are, due to the fact that they involve measurement and observation. Constraints are imposed by nature, they are physical laws. An example is that of trying to discover the color of an unused camera film. To do so, one has to take the film out of the roll. This is not sufficient, though, because one needs to have some illumination to see colors. As soon as we use light the film gets exposed and its color changes. Beside this simplifying example, physical principles let us conclude that in general:

sensing affects the state of the sensed entity.

One may be inclined to think that the above sentence is not true in general, for example seeing might seem to constitute a counterexample. When a human being sees an object, the eye absorbs photons coming from the sensed object, may they be reflected or emitted. This absorption amounts to the irreversible transformation from electromagnetic to chemical of the energy carried by these sub-atomic particles.

As the unfaithful Thomas story reminds us, human beings naturally tend to think that *truth is in seeing*, conferring vision a paramount role. Humans are inclined to identify the sensed object with the sensation of it. But, seeing is probably the most illusive of our senses. If we see a basketball player shooting at a game in the stadium, there must be one. But what if we see a very far

star? Is it there? Probably it imploded thousands of years ago, and we see the light emitted before the implosion.[5]

For the purposes of computer vision, not all physical facts sketched above have the same relevance; in our opinion, spatial coherence, presented next, is highly relevant.

3.1 The Spatial Coherence Constraint

Giving a definition of spatial coherence is not a straightforward task. To the best of our knowledge there is no such definition in the literature, where it is usually explained in terms of its effects. We do not attempt to give a definition here, rather we explain the constraint and show its relevance for spatial theories.

When an object in the real world emits (or reflects) an infinitely detailed and complex field, perception can handle this field only by reducing the incoming complexity by means of sizeable number of probes to locations. These probes are (by the indeterminacy of the scale of perception) hierarchically ordered in such a way that one probe is sensible to more detail, but over a smaller field of view and the higher ordered probe has less detail and a broader field of view.

This translates into the fact that the eye is sensible to certain modifications patterns but not to others. If one draws an arrow on a piece of paper, it does not matter if the hand was a little shaky, nor if the pen was defective and at some point more ink came out and at other less; any human being is able of recognizing it as an arrow. The observer is trying to capture the coherent structure of the field, thus the local thickness of the line can't affect the overall matching process.

Many successful theories have been devised in the field of spatial reasoning, for example in the field of diagrammatic reasoning: Euler circles or Venn diagrams [Ham95], [Gur98]. These theories do not consider the spatial coherence constraint, resulting in the impossibility of applying these spatial reasoning formalisms to any artificial vision system.

The dissertation "Languages of Perception" [Das98] of Mehdi Dastani, presents another interesting example of a nice theory of space hard to apply, due to the failure in satisfying the spatial coherence constraint. In his work, Dastani identifies languages for gestalt of visual patterns. The languages provide a compact and elegant way to group elements in patterns and to tie them together very naturally. The problem is that slightly modifying a pattern, maintaining its spatial coherence, results in any number of alternative new sentences describing the new pattern; whereas the human perception (the human 'sentence') of the pattern would be a single one.

Let us consider four balls placed at the same distance of one another interleaved with four boxes (see Pattern 1, in Figure 3.1). This pattern could be described, in the language defined by Dastani, by a couple composed of a ball and a box at distance δ iterated four times. If we now move one of the balls to the right (Pattern 2, in Figure 3.1) of an amount ϵ, the sentence describing the

[5] Another rich source of counterexamples to the principle of *truth is in seeing* is offered by optical illusions. For an interesting survey we refer the reader to [CG78].

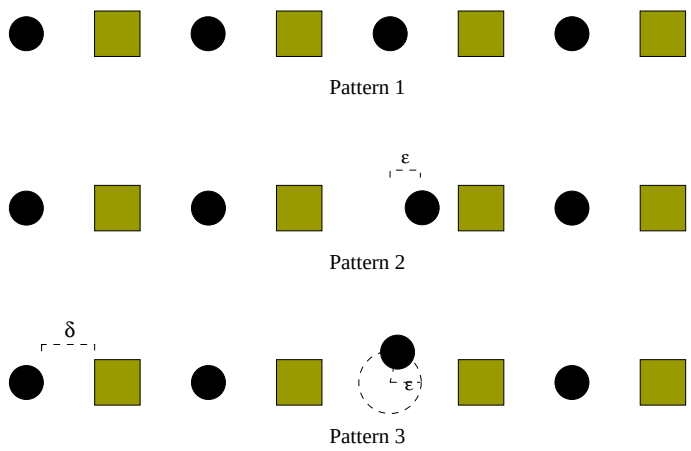

Figure 1: Patterns in space.

pattern will have to account for this difference. Namely, the moved element will be composed with the box to his right. If we now consider the same situation, but we move the ball in any other direction of the same amount ϵ, we get a different composition of the elements in the pattern. Spatial coherence imposes that in whatever direction the ball is moved of a fixed amount ϵ (Pattern 3, in Figure 3.1), the elements should be grouped in the same way, thus resulting in the same kind of pattern.

We believe that the work of Dastani is very valuable, especially when applied to the field of data visualization. His languages are very appropriate to describe the generation of visual patterns, to define spatial arrangements that carry specific intelligible meanings. What we argue is that those languages are not suitable to describe the way humans percept patterns in general. In other words, it is the title of the dissertation that is, in our opinion, misleading: we find "Languages of Patterns" more appropriate.

Luc Florack also writes about the coherence problem. In [Flo97], he identifies what he calls the "single and fundamental misconception" regarding image representation, i.e. the use of functions mapping spacetime regions onto subsets of the real numbers. He argues:

> "If we want to dispense with the abovementioned deficiencies, then we should *explicitly* account for:
>
> - the notion of spacetime as a *coherent* structure, and
> - the notion of *measurement*."[6]

The question of coherence is tied to that of neighborhood in a topology. Florack thinks that we cannot hope to simply apply topological concepts to space, but rather "enforce" a topology, which in turn can be achieved only a posteriori via the notion of observation.

[6] Italics in the original.

In the spatial reasoning community the problem is even more evident. Not only a suitable topology for space is not enforced, but often coherence on topologically defined neighborhoods is not considered at all, thus making any practical use of such formal theories impossible.

4 Possible Rendez-Vous: Image Bisimulations

We want now to probe further the mutual importance of computer vision and logic, and we especially want to hint which are possible rendez-vous.

The task is lifting existing IP technology to formally address the question of the semantic contents of images. We want to be able to extract semantic structures. Think for example of image retrieval systems. Current technology is either based on symbolic descriptions of images,[7] or on syntactic similarity measures performed by some set of IP functions (see [AAdR99] for a discussion). If we had automatically generated semantic descriptions of images, then we could have image retrieval systems correctly working on any set of images, enabling content based queries.

The first question is: how should these structures look like? We are inclined to select modal logic as the appropriate language to deal with low-level interpretation of images, for similar reasons for which modal logic is often employed to model time [vB83]. The spatial models we are looking for in this path should be similar to the models for most temporal logics (e.g., Kripke models, neighborhood models, topological models). Furthermore, we believe that the full computational power of first-order logic is unnecessary for many basic visual tasks.

The next question is: how do we process such structures? If our answer to the previous question is correct, then we aim at using the tools devised for temporal logics or straightforward extensions thereof. Let us consider the case of the image retrieval system. When each pair of two images has an associated model of its symbolic description, we can then restrict comparison to the matching of the two models. The comparison of models maybe on elementary equivalence, or on bisimilarity.[8]

Taking advantage of the parallel with temporal logics, we wonder about playing model comparison games. Wouldn't it be fun if we could play a kind of Ehrenfeucht-Fraïssé game[9] over images? Spoiler and Duplicator could be given two photographs from your collection of family trips. If Spoiler wins the game in one round, then you know that the pictures are very different, perhaps from two distinct trips. If Duplicator wins in, let's say, 10 rounds, then you know that both pictures were from the same trip and probably the same place on the same day: a picture of Niagara falls and one of you and your wife in front of

[7] These descriptions are attached to every single image by a human operator, thus all the semantic extraction and manipulation is performed outside the analyzing system.

[8] In his PhD dissertation [vB76], Johan used the word p-morphism to introduce a concept very close to that later named bisimilarity. The word bisimilarity was first used in the field of reactive systems.

[9] For a formal definition of these games see, for instance, [Doe96], while for an implementation and some example plays we refer to [AA99].

Figure 2: The map example.

the falls. If Duplicator can win an infinite game, then probably he has a copy of the same picture as Spoiler has or, more precisely, the two pictures have no perceivable differences from the point of view of the logic describing them.

We now want to take into full account the spatial coherence constraint. We think that also here modal logics can provide an answer. Intuitively, we want to be able to represent the underlying process occurring in the hierarchy of sensors from finer to coarser scale. We want to model the scale of the space in terms of a modal operator that brings us from a scale to another one. We have in mind something similar to the *microscopization* process proposed in [AV95]. In their work, Asher and Vieu introduce microscopization in terms of a predicate modal logic. Their main goal is giving semantics to natural language expressions of spatial regions. In addition, they aim at accounting for the difference between formal topology and a more intuitive type of topology, the one that some researchers think humans use in their everyday reasoning.[10]

Another interesting approach to model the change of scale is that presented in [vLvdD97]. In this setup, the process is referred to as *refinement* and it is modeled in terms of the inverse limit of a sequence of first order models. One of the goals of this work is a logic to explain the work of Marr [Mar83], work that only partially has passed the test of time, but some fundamental ideas are still valid.

For example, the idea of refinement can be one that helps us satisfy the spatial coherent constraint, thus it should be kept in our logic. Let us suppose given a description of a city in terms of a modal logic. The original description

[10] In the field of qualitative spatial reasoning this type of topology is known as mereotopology.

is in terms of main streets and neighborhoods. We now take two overlapping areas on the map and "zoom in" getting to a description in terms of blocks, smaller streets, and so on (Figure 2 should give an idea of the process). This zooming would be represented by our scale modal operator. The constraint on this zooming would be that the spatial properties perceivable on the general map, have to be also perceivable on the smaller ones, and be coherent with the general map. The question is whether this "also perceiving" should be identical or modulo a proper scale refinement condition.

Even thought we don't believe it possible to achieve a fine grained level of semantic description of images as the one necessary, for instance, to play Ehrenfeucht-Fraïssé like games over images (at least in the near feature), we do believe that finding suitable ways to represent semantic spatial content that do satisfy the spatial coherence constraint is a crucial question.

We will concentrate our future efforts on ways to satisfy such a constraint and on ways of representing and dealing with spatial models, believing that these are the first steps in the race to achieve artificial vision.

Acknowledgments We are thankful to Carlos Areces for comments on a previous version of this paper and to Massimo Canu for discussing some physical aspects presented here.

References

[AA99] A. Agostini and M. Aiello. Teaching via the web: A self-evaluation game using Java for learning logical equivalence. submitted, 1999.

[AAdR99] M. Aiello, C. Areces, and M. de Rijke. Spatial reasoning for image retrieval. submitted, 1999.

[AV95] N. Asher and L. Vieu. Toward a geometry of common sense: a semantics and a complete axiomatization of mereotopology. In *Proceedings of the IJCAI95*, pages 846–852. International Joint Conference on Artificial Intelligence, Morgan and Kaufmann, 1995.

[CG78] S. Coren and J. Girgus. *Seeing is deceiving: The psychology of visual illusions*. Hillsdale: Erlbaum, 1978.

[Das98] M. Dastani. *Languages of Perception*. PhD thesis, University of Amsterdam, 1998.

[Doe96] K. Doets. *Basic Model Theory*. CSLI, 1996.

[Flo97] L. Florack. *Image Structure*. Kluwer Academic Publishers, 1997.

[Gur98] C. Gurr. Theories of visual and diagrammatic reasoning: Foundational issues. In *Formalizing Reasoning with Diagrammatic and Visual Representations*. AAAI Fall Symposium, Orlando, Florida, 1998.

[Ham95] E. Hammer. *Logic and Visual Information*. CSLI and FoLLi, Stanford, 1995.

[Mar83] D. Marr. *Vision : A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman & Co, 1983.

[SKG98] A. Smeulders, M. Kersten, and T. Gevers. Crossing the divide between computer vision and DB in search for image databases. In *Visual Databases*, 1998.

[vB76] J. van Benthem. *Modal Correspondence Theory*. PhD thesis, University of Amsterdam, 1976.

[vB83] J. van Benthem. *The Logic of Time*. Kluwer, 1983.

[vLvdD97] M. van Lambalgen and J. van der Does. A logic of vision. Technical report, ILLC, 1997.