Noname manuscript No.
(will be inserted by the editor)

Towards Computational Cognitive Modeling of Mental Imagery

The Attention-Based Quantification Theory

Jan Frederik Sima and Christian Freksa

Received: date / Accepted: date

Abstract Mental imagery is the human ability to imagine and reason with visuo-spatial information. It is crucial for everyday tasks such as describing a route or remembering the form of objects. The so-called imagery debate has been centered around the question how mental imagery is realized, i.e., what structures and algorithms can plausibly explain and model mental imagery. There is, however, little progress on a coherent theory that can sufficiently cover the diversity of the empirical data. This article presents a new theory of mental imagery, which in contrast to other contemporary theories is formalized as a computational cognitive model. We will compare this theory to the contemporary theories using two representative phenomena of mental imagery. We will argue that formalized theories can advance the currently stagnant imagery debate.

Keywords Cognitive Modeling · Mental Imagery · Imagery Debate · Visuo-Spatial Representations

1 Introduction

This article presents a computational cognitive model of human mental imagery. Cognitive modeling can be understood as an approach to achieve *strong AI* (Searle, 1980) in the sense that cognitive models aim at the emulation of human intelligence; i.e., they do not only aim at the reproduction of the outcome of a task but in particular aim at the construction of internal structures and algorithms that directly correspond to those of the human cognitive apparatus. That is, a cognitive model is an approximation of the mental representations and processes of the human mind. In order to build a model of a given cognitive function a psychological theory of how the human mind realizes this function is necessary. A cognitive model is an implementation of a sufficiently formalized and detailed psychological theory. Cognitive models are evaluated by comparing their behavior, i.e., simulations, with the behavior of human subjects in different experimental settings. If a model can match human behavior and furthermore successfully predict human behavior for new experimental settings, there is reason to believe that the processes and representations of the model to some extent capture aspects of the human mind. The goal of cognitive modeling is the advancement of our understanding of the inner workings of the human mind. One ultimate prospect of such computational theories of human cognition is to tackle the hard problems of artificial intelligence, which seem trivial or natural for humans but remain impossible for computers so far, e.g., human-like conversation, imagination and creativity, and consciousness.

The presented cognitive model deals with visuo-spatial mental imagery. We experience this type of mental imagery whenever we visually imagine objects, scenes, persons, or routes. We have the experience of "seeing" something similar to visual perception without the object actually being there. Mental imagery is involved in fundamental aspects of human cognition, such as our ability to reason about spatial and visual problems. There has been a prominent debate about the mental structures and processes underlying mental imagery since the 1970s. This imagery debate used to essentially center around the question whether the mental representation underlying mental imagery is structured more like a picture or more like a description. The debate has not yet converged towards a conclusion, but, in fact, both main opponents see the status of the de-

Cognitive Systems, University of Bremen
Tel.: +421-218-64183
E-mail: {sima, freksa}@sfbtr8.informatik.uni-bremen.de

bate to be strongly in their favor (Kosslyn et al, 2006; Pylyshyn, 2002), while others propose both traditional theories to rely on fundamentally wrong assumptions (Thomas, 1999).

This article will give a brief overview of the contemporary theories of mental imagery and then introduce a new theory and its computational cognitive model. Using two important phenomena of mental imagery, we will compare the new theory with the contemporary theories regarding their explanatory and predictive power.

2 Theories of Mental Imagery

There are three main theories of mental imagery: the quasi-pictorial theory, the descriptive theory, and the enactive theory. The first two are the traditionally opposing accounts in the imagery debate. We will briefly review the three theories regarding their core assumptions.

Kosslyn (e.g., Kosslyn, 1994) has primarily shaped the **quasi-pictorial theory** and elaborated it in a framework containing several subsystems. A core assumption of the quasi-pictorial theory is the *existence of a depictive mental representation* in which mental images are *generated, inspected, and manipulated by processes also used in visual perception*. The mental image is depictive in the sense that spatial properties of what is imagined are represented by the spatial properties of the representation itself, e.g., the distance between two points is represented by distance within the mental representation. *Mental images convey their meaning via resemblence to what they represent.*

The **descriptive**, or **propositional**, **theory** is most prominently proposed by Pylyshyn (e.g., Pylyshyn, 2002) as a null-hypothesis to the quasi-pictorial theory. Essentially, it states that there is no need nor sufficient evidence for a special, i.e., depictive, mental representation of mental images. This means that the phenomena of mental imagery can be accounted for using *non-analogical, description-like mental representations*, which are assumed to underlie all high-level cognitive functions. The descriptive theory is extended with the concept of *tacit knowledge* (Pylyshyn, 1981). Tacit knowledge is the (subconscious) knowledge of what visual perception is like. It is used during mental imagery to *simulate the behavior of visual perception*, e.g., imagining a given stimulus means simulating what it would be like to visually perceive that stimulus.

The **enactive**, or **perceptual activity**, **theory** transfers concepts of active vision to mental imagery. Thomas (1999) sketches how enactive theory can explain mental imagery. Contrasting the other two theories, enactive theory rejects an explicit mental representation of mental images and instead proposes *implicit representations* called *schemata*. Schemata can be understood as subconscious processes that guide the recognition of an object during visual perception, e.g., which saccades are made to recognize a given object as that object. Mental imagery is explained by the execution of the respective schemata in absence of what is imagined, e.g., seeing a cat even though it is not there. In that sense, mental imagery is the *enactment of the visual perception of what is imagined*.

3 Attention-Based Quantification Theory

The attention-based quantification theory (ABQT) is a new theory of visuo-spatial mental imagery (previously described in Sima, 2011) that is implemented as a computational cognitive model. In order to understand how the theory proposes mental imagery to work, we have to take a look at the proposed relationship between visual perception and mental imagery.

3.1 Visual Perception as Abstraction - Mental Imagery as Concretization

The theory is based on the following assumptions about visual perception. Visual perception is understood as a process of abstraction. When we perceive a scene, we go through several attention shifts, e.g., saccades and micro-saccades, which result in the recognition of the visual features of the scene and its objects, e.g., lines, edges, and spatial relations between objects. These visual and spatial features are concrete in the sense that they are metric. For example, making a saccade along the edge of an object reveals the edge's coordinates in space. Given a set of recognized features, we can identify the object and its properties, e.g., qualitative size. Similarly, spatial relations are derived from a mapping of metric information, i.e., saccades between objects, onto qualitative spatial relations. When we later remember a scene from memory, we do not remember the exact metrics but the qualitative interpretation of that scene. Figure 1 shows a simple example of this abstraction in visual perception.

We can use qualitative information to reason. For example, we can decide which of two animals is larger if the associated qualitative sizes differ, but we need quantitative information if the qualitative sizes cannot be distinguished. Inferring that an elephant is larger than a mouse is faster than inferring whether a mouse is larger than a hamster (Kosslyn, 1980). We propose

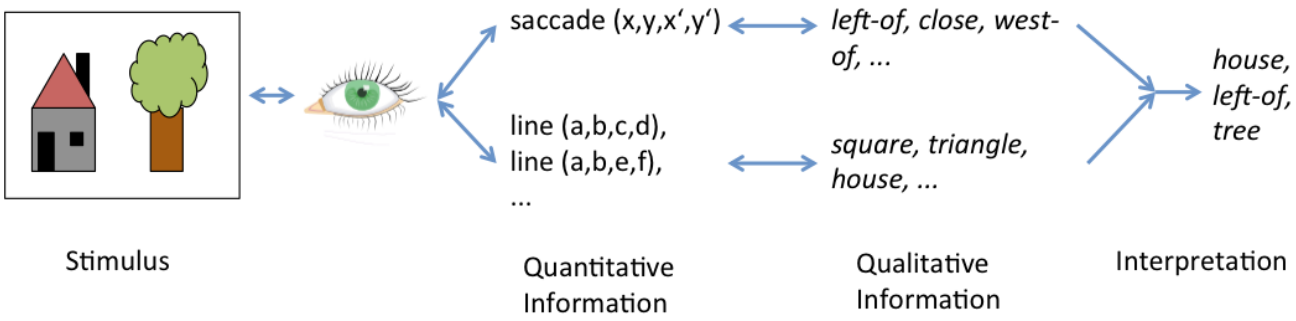

Fig. 1 Example of abstraction in visual perception. Attentions shifts, e.g., eye movements, lead to the recognition of objects and spatial relations in a scene. The initially quantitative information, e.g., visual features with coordinates, is abstracted to one qualitative interpretation.

that mental imagery is used in the latter case to make the concrete metrics of the respective shapes available to allow a quantitative comparison of the two animals. That is, mental imagery is the re-creation of one concrete instance of a scene based on its abstract description from long-term memory. Concluding, the purpose of mental imagery is to allow reasoning with concrete information.

The ABQT states that a mental image is based on abstract (qualitative) descriptions which are successively made concrete through the execution of attention shifts. These attention shifts are partially observable as spontaneous eye movements. The process of making abstract information concrete is called quantification. Quantification uses our implicit knowledge of visual perception. The mappings of attention shifts to spatial relations as well as to visual features and objects are used "inversely" during mental imagery. Attention shifts corresponding to a given shape are retrieved and the metrics of that shape are made conscious through the execution of the attention shifts. This process underlies the experience of "seeing" mental images.

3.2 Computational Cognitive Model

The computational cognitive model described in this section is an implementation of the attention-based quantification theory. The model consists of two long-term memory and two working memory components. There is a general long-term memory which stores explicit qualitative information and a visuo-spatial long-term memory which is specific to visual perception. The latter contains not consciously accessible, implicit knowledge about how to perceive objects and spatial relations during visual perception. There is a general working memory component that initially holds a qualitative description of a to-be-imagined scene and successively requests the quantification of its qualitative content. This quantification is realized by the other working memory component: the visuo-spatial attention window, which executes attention shifts and thereby makes implicit information explicit and available for further processing.

In the following, we describe the functionality of each component of the model. Figure 2 shows an example of how the components interact during the imagination of a scene.

The **long-term memory (LTM)** holds qualitative information. It is structured as a directed graph. The nodes represent scenes, objects, and parts of objects. The edges are part-of relations. Each node can contain properties, e.g., qualitative size of an object, and spatial relations to other objects. During mental imagery, the working memory (WM) queries the LTM for a given scene and retrieves the scene node and all its immediate child-nodes, which are the objects present in the scene. At a later point additional information, e.g., the parts of an object, i.e., its child-nodes, can be retrieved similarly when necessary.

The **working memory (WM)** contains qualitative information retrieved from the LTM as well as quantitative information retrieved from the visuo-spatial attention window (VSAW). It is structured as a directed graph. The nodes are temporarily extended with quantitative information. During mental imagery the WM successively requests quantification of the objects and spatial relations of the to-be-imagined scene. For this purpose the WM inputs the label of an object/spatial relation and if available additional qualitative and/or quantitative information to the visuo-spatial long-term memory (VS-LTM).

The **visuo-spatial long-term memory (VS-LTM)** holds information about a) how to look at a scene in order to recognize visual features which form familiar objects, and b) how to map eye movements to spatial relations. This information is used automatically during visual perception. During mental imagery this in-

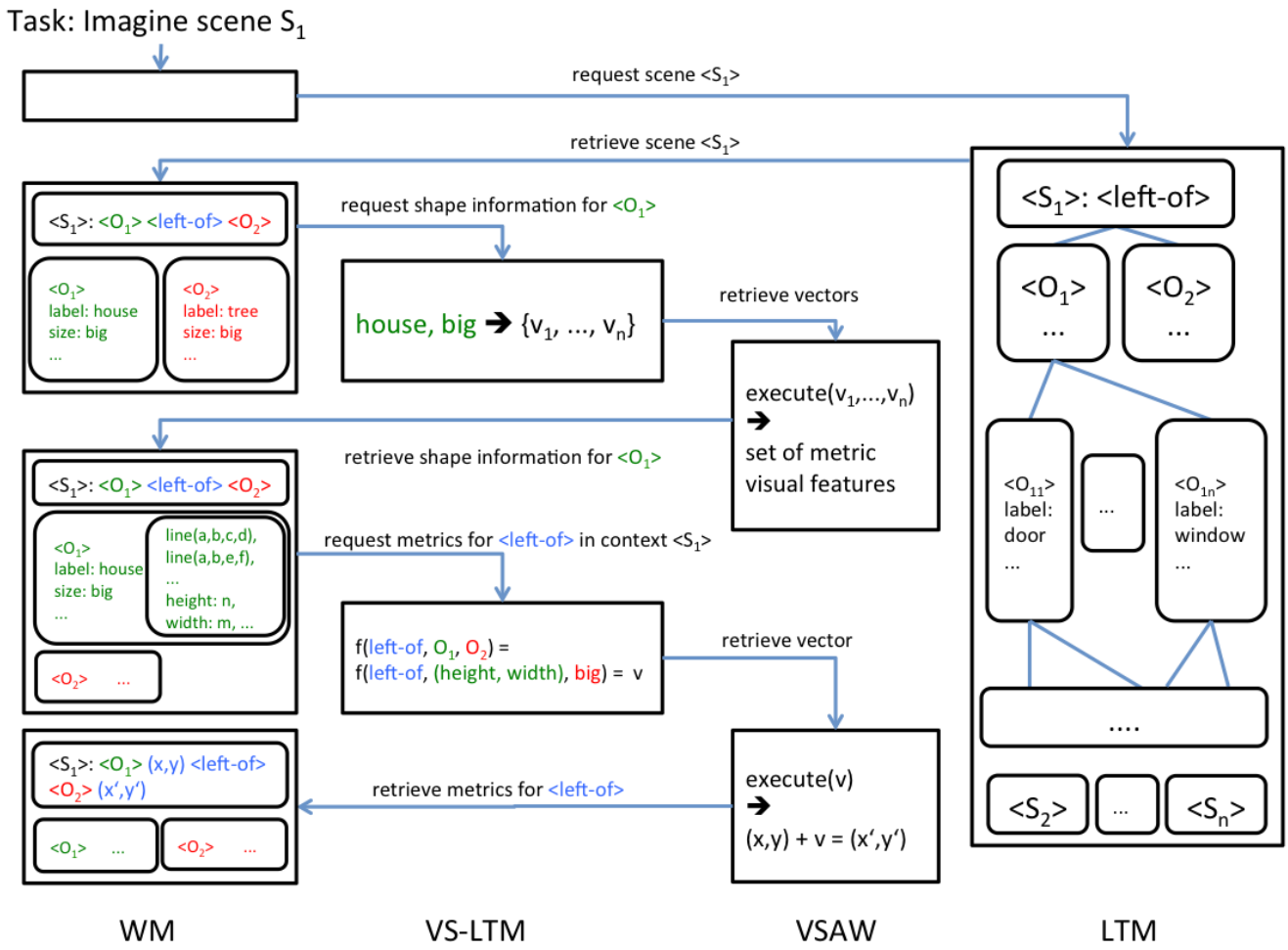

Fig. 2 Exemplary run of the model for a mental image of the stimulus from Figure 1. First, the shape information of *house* is quantified. Second, the spatial relation *left-of* is quantified. This assigns a location to *tree*. Further steps would be the quantification of the shape information of *tree* and if necessary a retrieval and quantification of further parts of *house* or *tree*.

formation mapping works "inversely": A given label of an object is mapped to a set of vectors. These vectors represent the attention shifts that are used to recognize the object. Similarly, the label of a spatial relation is mapped to a vector; this vector represents an attention shift associated with the spatial relation. The VS-LTM receives input from the WM and outputs to the VSAW.

The **visuo-spatial attention window (VSAW)** executes attention shifts during visual perception and mental imagery. During mental imagery the VSAW receives vectors as input from the VS-LTM and executes these as attention shifts. This process makes the implicit information of the VS-LTM explicit. The resulting quantitative information, e.g., metric visual features or a quantitative spatial relation, are returned to the WM. The VSAW is implemented by a pair of coordinates and a transformation function. The latter shifts the coordinates according to the vectors given as input. This transformation is a spatio-analogical process, i.e., shifting from one coordinate to another goes through all the coordinates in between. This is rooted in the assumption that the VSAW employs motor processes of visual perception, e.g., to execute eye movements, which are physically constrained in this way.

4 Comparison of the Theories

In this article, we will focus on two prominent phenomena of mental imagery: mental scanning and mental reinterpretation. These phenomena are particularly interesting as they are both complex phenomena that have been the subject of many empirical studies. Furthermore, they show both the diversity and the problems of the respective explanations of the contemporary theories.

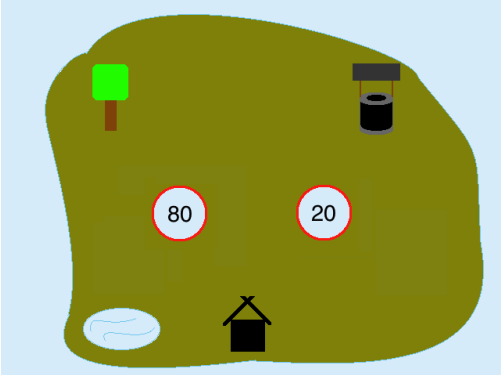

Fig. 3 Island stimulus similar to the one used in (Richman et al, 1979); the "80" and "20" signposts indicate the distances between the hut and the tree and between the hut and the water well, respectively. Note that these implied distances are obviously inconsistent with the actual metrics of the stimulus.

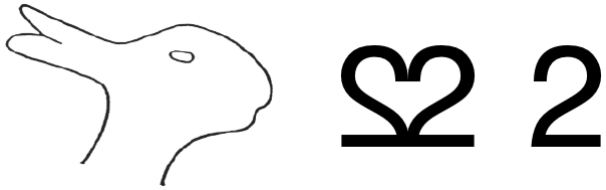

Fig. 4 Left: the duck-rabbit is hard to reinterpret mentally unless hints regarding the reference frame are given (Peterson et al, 1992). Right: most subjects can mentally reinterpret the right side of the heart-shaped stimulus as a "2" (Slezak, 1995)

4.1 Mental Scanning

Mental scanning is the process of shifting attention between two parts in a mental image. Many studies have shown that the time of this scanning process increases linearly with the distance between the two parts (e.g., Denis and Kosslyn, 1999). The so-called mental-scanning-effect is consistent with the quasi-pictorial theory's assumption of a depictive mental representation. Mental scanning is the successive shifting of attention on the mental image and subject to the inherent metrics of that image. The descriptive theory accounts for these results by posing that subjects use their tacit knowledge of what it would be like to visually scan the stimulus. The respective reaction times are simulated according to how long subjects believe it would take to visually scan the given distance. The enactive theory accounts for this phenomenon with its assumption that the execution of schemata during imagery is (at least partly) subject to the same constraints as during visual perception, i.e., scanning longer distances takes longer than shorter distances during perception. The ABQT explains the mental-scanning-effect similarly to the enactive theory as the VSAW is hypothesized to engage motor processes used in visual perception, e.g., eye movements.

It is, however, the special cases of mental scanning that reveal the complexity of the phenomenon. One such experiment is reported in (Richman et al, 1979): subjects studied a map similar to the one shown in Figure 3 and afterwards imagined the island as a mental image. When scanning between the different entities, there was a significant effect of the distance signposts on the scanning time, i.e., the route with the distance signpost *80* resulted in longer scanning times than that with the distance signpost *20* despite the fact that these implied distances are inconsistent with the actual stimulus, in which both distances are equal.

For the **quasi-pictorial theory** to be able to explain this result either the depictive representation would have to be distorted compared to a normal mental scanning task or the speed of the inspection process would have to be adjusted accordingly. It is unclear how any of the options would be realized.

It is also unclear how the **enactive theory** could explain such results given that an enactment of the visual perception of the island should not show these differences in scanning time.

The **descriptive theory** interprets such results as evidence for its tacit knowledge hypothesis and against the quasi-picture hypothesis. The result allegedly shows that the mental-scanning-effect is cognitively penetrable[1], i.e., it is not a result of the (depictive) format of an underlying mental representation but can be manipulated by a subject's belief, i.e., a subject's tacit knowledge. The actual mechanisms are, however, not explained.

The explanation offered by the model of the **ABQT** can be understood as a hypothesis about how cognitive penetration of mental imagery tasks comes about in general. The distance signposts are encoded in the qualitative spatial relations, i.e., the *20* signpost could encoded as, for example, *water well - upper right of, near - hut*. The quantification of these spatial relations is realized through the execution of different vectors by the VSAW, i.e., *upper left of, far* is mapped to a longer vector than *upper right of, near* by the VS-LTM. As the VSAW's execution of a vector takes more time the longer the vector is, the model shows the observed difference in reaction times.

4.2 Mental Reinterpretation

Mental reinterpretation is the discovery of a second meaning of an ambiguous stimulus using only a mental image of that stimulus. The literature shows that some stimuli are relatively easy to reinterpret mentally while others are relatively hard unless subjects are given hints about the necessary reference frame transformations. Figure 4 shows one example of each type of ambiguous stimuli. The "mirrored 2" represents a class of relatively simple stimuli made up of basic geometric forms and alphanumeric characters (mostly used in Finke et al, 1989). The successful reinterpretation of the duck-rabbit (and similar stimuli) usually requires either implicit hints, i.e., examples of ambiguous stimuli that require the same reference frame transformation, or explicit hints, i.e., "the back could be the front of another animal" (e.g., Peterson et al, 1992). The challenge a theory of mental imagery has to face is why and how some stimuli are reinterpretable mentally while others require certain conditions.

Quasi-pictorial theory argues that, for example, the duck-rabbit is hard to reinterpret mentally because the constituent parts of the mental image keep fading so fast that a sufficiently complete "image" cannot be inspected and therefore not be reinterpreted (Kosslyn, 1994). This explanation can account for the difference between the duck-rabbit and less complex stimuli like the "mirrored 2", but it does not explain how and why the reinterpretation of the duck-rabbit improves with implicit and explicit hints. Furthermore, such an initially unexpected rapid fading of the parts of mental images makes one wonder about the consequences for other mental imagery tasks with similarly complex stimuli, e.g., mental scanning.

From the perspective of the **descriptive theory** the difficulty of mental reinterpretation for stimuli that are easily reinterpreted during visual perception weakens the hypothesis of a depictive mental representation and thus supports the null-hypothesis of a description-like mental representation. The relatively easy mental reinterpretation of more trivial stimuli can be attributed to inference processes based only on descriptions and not depictive information (Pylyshyn, 2003). No concrete statements about how hints facilitate mental reinterpretation or how exactly trivial stimuli are reinterpreted are made.

The **enactive theory** proposes the difficulty in reinterpreting the duck-rabbit stimulus to result from the fact that subjects only perceive the initial stimulus as, for example, a rabbit. During mental imagery they subsequently enact the visual perception of the rabbit, i.e., employing schemata leading to the interpretation of the stimulus as a rabbit, inhibiting the option of "seeing" it as a duck (Thomas, 1999). This explanation omits the mechanisms of reference frame hints and only briefly suggests that less complex stimuli like the "mirrored 2" might be dealt with differently due to a specific familiarity with alphanumeric characters.

The **ABQT** poses that the difference in how easy or hard stimuli can be reinterpreted mentally lies in their representation in working memory. More complex stimuli, like the duck-rabbit, consist of several distinct parts (nodes), e.g., *nose*, *ears*, with semantic labels and spatial relations between them. In contrast, the "mirrored 2" stimulus can be represented by one or two parts (nodes) and does not itself have a clear semantic label. These two properties, i.e., number of parts and meaningful labels, determine the difficulty of mental reinterpretation. During mental imagery a stimulus is quantified part by part. When the quantitative information of a part is available there is the possibility of a (re-)interpretation just as in visual perception (see Figure 1). This (re-)interpretation is constrained by the current context. During mental imagery the WM already holds a qualitative interpretation of the whole object whereas during visual perception this final interpretation is just then created. The reinterpretation of the "mirrored 2" is easier because there is no explicit semantic label of what the initial stimulus is and few or no other parts have to be considered in the interpretation. Note that a subset of the visual features of the "mirrored 2" corresponds to the visual features of *2*. In case of the duck-rabbit the quantified shape information of, for example, *ears* could in principle be interpreted as *beak*. But that part is already labeled as *ears* in WM and, furthermore, *beak* does not fit with the other parts of a rabbit and also does not fit at the back (as specified by the current spatial relations) of any animal. The implicit or explicit hints that improve successful reinterpretation of the duck-rabbit give precisely the information how to alter the spatial relations of the parts in WM or to abandon the current interpretation of the parts altogether ("the back is the front of another animal") and thus facilitate the necessary reinterpretation of each single part.

5 The Future of the Imagery Debate

In conclusion, we want to argue that the lack of progress towards a solution of the imagery debate is rooted in the lack of formalization of the contemporary theories. The insufficient formalization results in a detrimental vagueness inherent even to fundamental assumptions,

[1] Cognitive Penetration is explained, for example, in (Pylyshyn, 2002, p. 161)

for example, the undefined nature of a quasi-picture or the open issue of how tacit knowledge is stored and employed. Generalizing the problems of the contemporary theories with respect to the discussed mental imagery phenomena, we identify three major shortcomings of not adequately formalized theories: 1) vague conceptions are often not testable and therefore cannot be refuted empirically; 2) given vague core assumptions it is possible to generate nearly arbitrary additional hypotheses to explain new (otherwise possibly contradicting) empirical data; and 3) individual phenomena can be accounted for with specific assumptions, whose consequences for the explanation of other phenomena and the overall theory remain unclear.

Computationally implementing psychological theories as cognitive models enforces a level of detail that abolishes the mentioned shortcomings. An implemented theory of mental imagery can develop transparently and achieve further progress by driving empirical research with its more concrete assumptions and predictions. The empirical testing can in turn suggest concrete adjustments and refinements of the model. For example, as elaborated above, the ABQT was able to provide more detailed accounts for the discussed phenomena than the contemporary theories. Furthermore, the explanations lead to clear predictions, e.g., the criteria posed for the difficulty of reinterpretation, which can be tested in future empirical research to either confirm the theory or suggest adjustments. We therefore believe that the paradigm of cognitive modeling has the power of getting the so far divergent imagery debate out of its deadlock. The presented model of the attention-based quantification theory is a first step towards this goal.

Acknowledgements This paper presents work done in the project R1-[ImageSpace] of the Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition. Funding by the German Research Foundation (DFG) is gratefully acknowledged.

References

Denis M, Kosslyn S (1999) Scanning visual mental images: A window on the mind. Cahiers Psychologiques Cognitives 18:409–465

Finke RA, Pinker S, Farah MJ (1989) Reinterpreting visual patterns in mental imagery. Cognitive Science 13:51–78

Kosslyn SM (1980) Image and Mind. Harvard University Press, Cambridge, MA

Kosslyn SM (1994) Image and brain: The resolution of the imagery debate. The MIT Press, Cambridge, MA

Kosslyn SM, Thompson WL, Ganis G (2006) The Case for Mental Imagery. Oxford University Press, New York

Peterson MA, Kihlstrom JF, Rose PM, L M (1992) Mental images can be ambiguous: Reconstruals and reference-frame reversals. Memory and Cognition 20(2):107–123

Pylyshyn ZW (1981) The imagery debate: Analogue media versus tacit knowledge. Psychological Review 88:16–45

Pylyshyn ZW (2002) Mental imagery: In search of a theory. Behavioral and Brain Sciences 25(2):157–238

Pylyshyn ZW (2003) Seeing and visualizing: It's not what you think. MIT Press, Cambridge, MA

Richman CL, Mitchell DB, Reznick JS (1979) Mental travel: Some reservations. Journal of Experimental Psychology: Human Perception and Performance 5(1):13–18

Searle JR (1980) Minds, brains, and programs. Behavioral and Brain Sciences 3:417–424

Sima JF (2011) The nature of mental images – An integrative computational theory. In: Carlson L, Hoelscher C, Shipley T (eds) Proceedings of the 33rd Annual Conference of the Cognitive Science Society, Cognitive Science Society, Austin, TX, pp 2878–2883

Slezak P (1995) The 'philosophical' case against visual imagery. In: Slezak P, Caelli T (eds) Perspective on cognitive science: Theories, experiments, and foundations, Ablex, Norwood, NJ, pp 237–271

Thomas NJT (1999) Are theories of imagery theories of imagination? An active perception approach to conscious mental content. Cognitive Science 23:207–245

Fig. 5 Jan Frederik Sima is a researcher at the Cognitive Systems working group at the University of Bremen. He works in the project R1-[ImageSpace] of the Spatial Cognition Research Center SFB/TR 8. He studied computer science and political science at the Technical University of Darmstadt.

Fig. 6 Christian Freksa is a full professor at the University of Bremen and Head of the Cognitive Systems research group. His research interests are in the area of spatial cognition and spatio-temporal reasoning. Christian Freksa is the coordinator of the Spatial Cognition Research Center SFB/TR 8 and a Fellow of the European Coordinating Committee for Artificial Intelligence (ECCAI).