A Computational Account of Complex Mental Image Construction

Jan Frederik Sima (sima@sfbtr8.uni-bremen.de)
SFB/TR 8, Universität Bremen, Germany

Abstract

This paper presents a computational cognitive model of the construction process of complex, i.e., multi-part, visual mental images. The model is integrated into the cognitive architecture Casimir. The construction process is realized by the interplay of a spatial working memory structure and a passive quasi-pictorial visual representation. Both structures are successively build up on demand from long-term memory. The correct placement of new parts is guided by the inspection of the visual representation. The model has two main advantages: 1) it is an explicitly cognitive computational model that implements the two-fold structure of a spatial and a visual working memory representation and 2) it introduces an attention window structure in such a way that allows for direct predictions of eye movements during mental imagery processes. We discuss predictions and explanations offered by model.

Keywords: Cognitive Modeling; Visual Mental Imagery; Visual and Spatial Representations; Analogical Representations

Introduction

The experience of visual mental imagery is a well-known and widely studied phenomenon. For example, many people report to actively use mental imagery for common visuo-spatial tasks, such as planning a route. Additionally, imagery plays an important role in a number of diverse domains such as diagrammatic problem solving and creativity (e.g., Hegarty, 2004). Furthermore, the general efficiency and usefulness of a visual or quasi-pictorial representation compared to a purely symbolical, i.e., non-analogical, representation for several reasoning domains has been shown and argued for extensively from an artificial intelligence point of view (e.g., Chandrasekaran, Kurup, Banerjee, Josephson, & Winkler, 2004).

Almost all computational accounts of visual mental imagery that have emerged since Kosslyn's computational cognitive model (Kosslyn, 1980) thirty years ago were not designed as cognitively plausible accounts of human imagery processes, but adopted single findings, e.g., most prominently the existence and distinction of two, one spatial and one visual, representations involved in mental imagery (see for example Glasgow & Papadias, 1992). There has been work to extend well-established cognitive architectures, e.g., ACT-R and Soar, with the functionality of visual mental imagery and even though these accounts provided valuable insights, for example regarding the structural integration of imagery into an architecture, they remained on a conceptual level (Gunzelmann & Lyon, 2007) or were explicitly not designed as cognitively plausible models (Lathrop, 2008).

As Kosslyn's computational model (Kosslyn, 1980) is the most relevant and also closest in its approach to our model, it is worthwhile to make the major differences clear. First off, it is to note, that Kosslyn (1994) himself significantly altered his theory of mental imagery in the light of new empirical and neuroscientifc data. His new and very extensive conceptual model has, however, never been implemented. In contrast to his implemented model, we employ two working memory structures: the visual one roughly corresponds to Kosslyn's visual buffer, the other spatial one has no counterpart in his model. Another important difference is the existence of an attention window in our model, which implements the selective attention on the content of a mental images as well as the multi-scale property of the visual representation.

The aim of the presented model is to offer a plausible explanation of how complex visual mental images are constructed from long-term memory. The computational implementation allows the identification of open empirical issues as well as new predictions regarding the involved processes in mental imagery. By employing two working memory structures of different abstraction, we offer a straightforward account for the findings that suggest two distinct kinds of imagery (Levine, Warach, & Farah, 1985; Farah & Hammond, 1988) and also shed new light on the question of many imagery phenomena such as mental image reinterpretation. The implemented attention window of the model allows us to directly link attention shifts in the visual representation to eye movements made by subjects in imagery experiments and thus offers a new method of evaluation for models and theories of imagery.

The model is designed within the framework of Casimir (Barkowsky, 2007), a cognitive architecture for spatial knowledge processing with analogical representations.

In the following sections, we will describe the design of the model's representation structures and processes and how those are derived from general assumptions of human cognition as well as from empirical data on several related imagery phenomena.

The Model of Image Construction

To begin, we define the domain the model is applied to. When referring to mental images, we always mean consciously experienced visual mental images. The model is constrained to mental images that are generated from information that is retrieved from long-term memory with the absence of any other visual input, e.g., visual perception. The mental images we deal with are labeled "complex" in the sense that the visualized concepts consist of several parts. For example, the concept *house* consists of a main block, a roof, a door and a window; further, the parts themselves may have subparts, e.g., chimney is a part of the concept *roof*. The mental images are constructed so that they are "seen" from an egocentric perspective similar to an actual visual percept.

Figure 1 shows the basic components and interactions of the model. We have modeled the spatial and visual representations as well as the processes involved in the construction of a mental image. The long-term memory component

is an existing part of the cognitive architecture Casimir (see Schultheis, Barkowsky, & Bertel, 2006, for details).

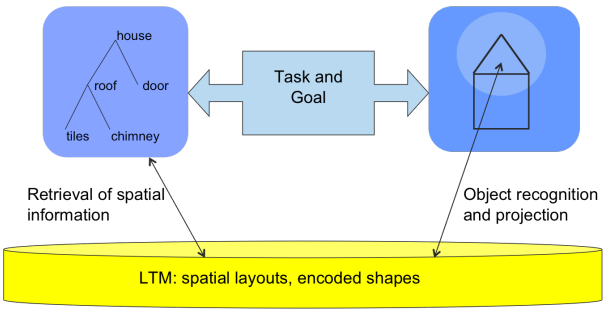

Figure 1: Representations and Processes of the Model. The visual representation serves as an extension of the spatial representation. Shapes are projected into the visual representation according to the spatial layout stored in the spatial representation.

Basic Design Constraints

In this section we will briefly elaborate the theoretical background upon which the general design decisions of the model are based. For this purpose, we describe the basic assumptions that the model makes about visual mental imagery and working memory in general.

Parsimony. The model is generally designed to keep the processes and corresponding representation structures as parsimonious as possible. The model's workflow is designed so that it works strictly on demand. This means, that each transfer and transformation of information between long-term memory, the spatial representation and the visual representation is only triggered when demanded by the current task. Accordingly, concepts can be visualized at different levels of granularity and enriched with more details when necessary.

Analogical representation structures. The model is based on the main assumptions of what is often labeled the quasi-pictorial theory of mental imagery, see (Kosslyn, 1994) for its most popular representative. That is, the structure or structures, which the experience of visual mental images relies on, at least partly represent/preserve the spatial properties of an actual image/the actual visual percept in an analogical format. Given the existing empirical support, there is widespread agreement on this hypothesis (e.g., Finke, 1989). As the visual representation in the model actually depicts shape it is apparently analogical, but also the spatial representation has an analogical format as it preserves the part-of relation of complex entities in its structure.

Distinction between visual and spatial knowledge processing. Within the model visual and non-visual information is distinguished on different but interdependent levels: 1) the model employs two working memory structures, 2) visual and non-visual information is retrieved separately by separate subprocesses from long-term memory.

Building upon the findings that the two cortical visual pathways first identified by Ungerleider and Mishkin (1982) can also be distinguished in human visuo-spatial working memory (Courtney, Ungerleider, Keil, & Haxby, 1996), it has been argued that two representations involved in imagery can be functionally and neurologically dissociated (Levine et al., 1985). This conclusion is based on studies with brain-damaged patients, who were able to perform normally on some imagery task but were impaired on other imagery tasks. These two groups of imagery tasks corresponded to what is usually considered to be visual imagery tasks and spatial imagery tasks respectively (Farah & Hammond, 1988).

As evident in figure 1 we assume two information pathways which together give rise to complex images in the visual representation. On the one hand, processes associated with the ventral pathway are responsible for the processing of shape information, i.e., the recognition of shape and the retrieval and projection of shape information from long-term memory into the visual representation during imagery. On the other hand, the spatial representation is associated with the dorsal pathway which processes the spatial layout of an entity or scene.

Besides fulfilling all other structural requirements for models of mental imagery as identified by Bertel, Barkowsky, Engel, and Freksa (2006), our model specifically fits into their category of hybrid models, as two representations of different qualitative structure are combined. They proposed that a computational cognitive model of mental imagery needs to have a hybrid structure in order to plausibly capture the "hybrid, exhibiting both visual and propositional traits" (Bertel et al., 2006) nature of mental images.

Evidence for a dedicated non-visual working memory structure involved in visual perception, has led to approaches (e.g., Nestor & Kokinov, 2004), which, similar our model, employ a visual and a non-visual working memory structure in this respective domain.

Components and their Interaction

Following, we will describe the structure of both the spatial and the visual representation in more detail as well as the interaction between them.

The visual representation is implemented as a graphics window, in which geometric shapes are drawn. The circular attention window determines which parts of the representation are currently attended to and can be processed. The attention window is defined by its position and by its resolution. The higher the resolution, the smaller the extent of the attention window and thus only a smaller part of the visual representation is accessible for inspection. Furthermore, the resolution also determines what contents of the visual representation are "visible", i.e., can be processed, depending on the size of the visualized shape. For example, small parts or details such as texture are only accessible if the resolution is high, whereas bigger parts are also visible at a low resolu-

tion. The attention window implements two concepts: 1) the selective processing of visual information and 2) the scale-resolution trade-off in the inspection of mental images, which goes along with the multi-scale property of the topographically organized areas of the visual cortex (Kosslyn, 1994).

The spatial representation contains the minimal necessary spatial layout information of a concept. For the concept *house* the minimal layout consists of a location[1], orientation and size of the basic shape of *house* as it is visualized or to be visualized in the visual representation. Note that these parameters can be set by the task, e.g., "Imagine a small house, that is tilted 90 degrees clockwise", but lacking any of those demands, the parameters will be set by associations from long-term memory. The spatial representation does not include the shape or any further information about the shape other than the rough size it is (to be) visualized in. The minimal layout further includes the direct and most strongly associated parts of the concept *house* as identifiers, their spatial relations to the basic part, e.g., "on the left top of", as well as their relative size compared to the basic part of *house*.

The relative size of a part is important to determine if and when it is visualized in case an elaborate image of the current concept is demanded, i.e., the bigger the part is relative to the basic shape[2] of a concept, the earlier it will be visualized. That is, if a detailed image of *house* is requested, *roof* will be visualized first, followed by *door* and *window*. We assume that this size-depended sequence might change if one particular part has a very strong association with the super concept, but as a default the relative size is assumed to determine the sequence. This is a consequence of the nature of the attention window and we further elaborate on this aspect below.

If a new part, e.g., the door of the house, should be visualized, the concept *door* is retrieved from long-term memory and extends the spatial representation. This means that it now includes information about *door*; orientation, size and location are in this case determined by the super concept *house*, e.g., if we imagine a small 90 degrees tilted house, all its parts and subparts will by default also have these properties. Parts of *door* are now also consciously available. The retrieval of new information is context-dependent as it is affected by the current content of the spatial and visual representation, that is, in particular the super concept, e.g., the model would produce a different mental image of a window by itself than of a window as part of a house.

Interaction between components. There is a hierarchical structure between the long-term memory, the spatial representation and the visual representation, that is, information is retrieved and transformed from long-term memory first into the spatial representation and parts of these informations are transferred on demand into the visual representation, where the resulting shape is visualized. As evident in figure 1, there is a direct connection between encoded shapes in long-term memory and the visual representation, but this projection process is triggered only if parts of the spatial representation need to be visually accessed. The represented information on these three levels differs quantitatively as well as qualitatively: 1) there is information available in the spatial representation which is not visualized, i.e., not represented, in the visual representation; similarly the information in the spatial representation is only a fraction of what is available in long-term memory; 2) furthermore, only the visual representation explicitly contains visual information, such as shape or texture, which by themselves lack semantics (which are contained in the spatial representation). Additionally, this hierarchical structure implies that certain tasks, which do not depend on visual information can be solved solely on the level of the spatial representation and do not have to use the visual representation.

The Image Construction Process

In order to describe the construction process of a multi-part visual mental image in the model, we will go through the individual steps taken to build an image.

- The model is given the command to imagine the concept *house*.
- The *spatial representation* (SR) queries the *long-term memory* (LTM) for the minimal spatial representation of *house*. As no further context is specified, a default location *L*, orientation *O* and size *S* are used for the query.
- The *attention window* (AW) is shifted to location *L* and its resolution adjusted to fit the size *S*.
- The *visual representation* (VR) retrieves the basic shape of *house* with the given size *S* and orientation *O* from LTM and it is visualized at the center of the AW.
- The SR is queried for direct parts of *house* that are of the same relative size as *house* and finds *roof*. It will automatically be visualized given the current resolution of the AW.
- The shape of *house* in the VR is inspected to find the coordinates where to place *roof* according to the given qualitative spatial relation between *roof* and *house* from the SR.
- The AW to the determined location.
- The SR retrieves further spatial information about *roof*; this information includes parts of *roof* and will allow for a later visualization of parts of *roof*.
- The shape of *roof* is retrieved from LTM with size *S* and orientation *O*, which are both inherited from the parental concept *house*. The shape is projected into the center of the AW.
- The SR does not find any other direct parts of *house* with a relative size that would allow visualization given the current resolution. Thus the model stops.

The above process sequence builds the minimal image of the concept *house*. Further parts such as *door* or *window* are

[1] Location within the visual representation.

[2] Following a similar principle the basic shape or main part of a concept figures to be the bigger than any of its parts.

not added, even though their existence, relative size and spatial relation are "known", i.e., are consciously available in the SR. The model always builds minimal images unless the task demands further details to be added.

Lets look at an excerpt of the construction process for a detailed image of *house*. We assume the state of the model to be the last described state of the previous process sequence, i.e., the basic shape of *house* and *roof* are visualized.

- When a detailed image is requested all directly related parts to *house* are visualized in the order of their relative size. The SR finds *door* as a part of *house*.
- The basic shape of *house* in the VR is inspected and the appropriate docking coordinates for *door* are calculated and the AW is shifted to this position.
- As the relative size of *door* is smaller than that of *house* the resolution of the AW is adjusted, i.e., higher resolution, lesser extent.
- The SR retrieves spatial information about *door*.
- The shape of *door* is retrieved and projected at the position of the AW in the VR.
- This process goes on similarly for all direct parts of *house*.

Explanations and Predictions

The model makes some novel assumptions which offer new and concrete explanations of common imagery phenomena and also lead to precise predictions about human behavior during mental imagery. We will briefly look at how the model is able to account for those common phenomena of mental imagery. We have not yet started to fit concrete empirical data, but the structure of the overall model and it's individual representations and processes strongly suggests that the model will at least reproduce the qualitative trends of the following phenomena. In the following, we will not cite single studies for each phenomenon, but we rather refer the reader to Kosslyn, Thompson, and Ganis (2006) for an overview of the mentioned studies.

Image generation. Empirical studies suggest, that the construction time of a mental image of a scene or object directly depends on the number of parts and the level of detail. The model offers a trivial and straightforward explanation, as it generates mental images piece by piece. What is more interesting and novel is the proposed sequence in which parts are added and we further discuss this point below.

Image scanning. Several different studies suggest that the time taken to mentally scan from one point of a mental image to another is proportional to the imagined distance between these points. The attention window of our model is shifted gradually over the visual representation to the respective portion of the visual representation that needs to be processed. Therefore again, the model provides a straightforward account of this phenomenon.

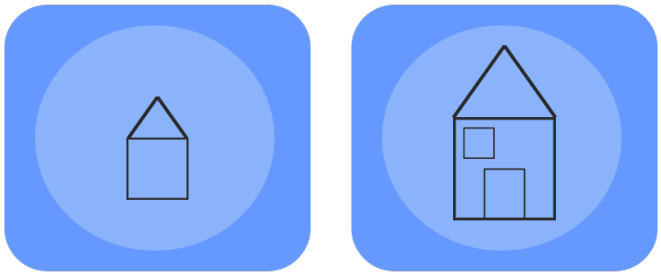

Figure 2: Example of Zooming. Resolution of the attention window is low and therefore only big (size==3) and medium (size==2) sized parts are visualized. Left side: the main shape of the concept *house* is imagined in medium size (size==2). The shape of roof is also visualized as it is of the same relative size (size of *house* plus the relative size of *roof*, i.e., $2-0=2$). *Door* and *window* have a small size (size of *house* plus relative size of *door*, i.e., size==$(2-1=1)$ and are therefore not visible given the current resolution.
Right side: The size of *house* was set to big (size==3) and therefore the size of *door* and *window* is now medium (size==2). Thus, they are now visualized.

Zooming. Zooming in or out of a mental image is realized by altering the size parameter of a concept or a part of the concept in the spatial representation. This parameter is used to determine the extent of the respective shape when it is projected onto the visual representation. Furthermore, if the size parameter is altered for a concept in the spatial representation and it is therefore re-visualized with a now bigger or smaller shape, this has automatic consequences for the visualization of the parts of this concept. The spatial representation stores a concept's parts with their relative size compared to the basic shape of the concept. This relative size again determines whether and when a part is also visualized in the visual representation given the current level of resolution of the attention window (see figure 2 for a visual example). The empirical findings regarding zooming in mental images express that it will take subjects more time to find a part of an imagined object if it is initially imagined at a small size than when it is imagined at a bigger size. These findings can potentially be explained in two ways by the model: 1) subjects employ a zooming process as described above or 2) the resolution of the attention window is increased which will also make some smaller parts of the image "visible".[3] Both accounts would results in increased processing time and thus qualitatively match the empirical results.

Image organization and reinterpretation. There is evidence that mental images have an underlying organization. It has for example been found, that the way presented stimuli are described, e.g., the star of david as either two overlapping triangles or as an hexagon with six small triangles, affects the

[3]For very small parts increasing the resolution will not work and zooming will be necessary.

way subjects later recreate this image mentally. That is, on the one hand, image generation takes longer when the image consists of more parts and on the other hand recognition of patterns that are congruent with the organization of the image is faster than for other valid patterns. A related phenomenon is the difficulty of mental image reinterpretation. That is, it is very difficult for subjects to reinterpret an ambiguous picture as a mental image, if that picture was preciously learned realizing only one of its meanings. Whereas, it is much easier to find the second meaning during normal visual perception of the same ambiguous picture. Both of these types of findings point towards the same direction of mental images being more than just a mental depiction of visual information but including semantics and depending on a more abstract structure and organization underlying the depictive structure. The two-fold structure of our model provides just that. As the spatial layout held in the spatial representation is used to build up the mental image in the visual representation, it is apparent that this consciously available organization affects how the content of the visual representation is inspected as well as interpreted. In order to successfully reinterpret an ambiguous image during mental imagery, the content of the spatial representation would have to be discarded, because even though the content of the visual representation might the be so that it depicts both meanings, the individual parts would need to be linked to different concepts. Furthermore, the retrieval of shape information from long-term memory is context-dependend regarding the currently held concept in the spatial representation. This means that a retrieved shape and its properties are affected by the concept it is linked to in the spatial representation; especially by background knowledge about a concept. For example, the mouth of the rabbit in the famous duck-rabbit image (see figure 3) might not be recalled when subjects are imagining a duck, because this visual feature is irrelevant for the shape of the back of a duck's head.

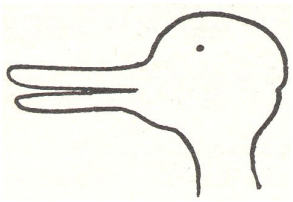

Figure 3: Ambiguous Duck-Rabbit Image

Predictions of the Model There are three main predictions we can draw from the model: 1) internal attention shifts are functional for the construction of complex mental images and are reflected by eye movements, 2) the sequence in which parts are added to a complex mental image is affected by the relative size of the parts, and 3) the visual representation is used only when demanded by the task.

1) Whenever a new part of an image is visualized in the model, the attention window is adjusted in its location and its resolution. That is, it is shifted to the location the new part will be visualized at. There are several studies (e.g., Johansson, Holsanova, & Holmqvist, 2005) that have shown a close correlation between eye movements and the currently processed contents of a visual image. The model implies that eye movements are linked to the shifts of the attention window during mental imagery. Furthermore, these attention shifts are functional to the process of mental image construction.

2) The construction process proposed by the model differs from previous assumptions about the sequence in which parts are added to form a mental image. A common default assumption seems to be that the sequence of parts is determined by the strength of association of the part with the main concept. Furthermore, this is often combined with the idea of choosing that part next, which yields the highest identification value for the concept. This idea stems from an analogy to top-down-hypothesis testing in object recogntion (see Kosslyn, 1994). In contrast, our model predicts a very different sequence for image construction, which is a direct consequence of the implementation of the attention window. The attention window has different scales of resolution, which determine whether a part is visualized and also whether a visualized part is accessible. That is, with the initial low resolution only big parts can be visualized and processed, whereas with a high resolution also smaller parts, i.e., details, are "visible". The model will according to its principle of parsimony not change its resolution, i.e., go into more detail, unless it is necessary. This means, that direct parts of the concept are visualized first when this is possible without a change of resolution, i.e., the ones that are closest in relative size to the concept's main shape.

3) Lastly, the hierarchical structure of the model allows for an on demand usage of the visual representation. That is, if visual information, like the exact shape, is not necessary to fulfill a task, the processing will remain on the level of the spatial representation. This concept fits nicely with the work of Sima, Lindner, Schultheis, and Barkowsky (2010), who found that the same spatial reasoning task is solved by either using mental imagery or by using a more abstract representation, e.g., mental models, depending on whether the instruction demands imagery or not.

Conclusion and further work

We have presented a computational cognitive model of human complex mental image construction and elaborated on the underlying assumptions as well as the predictions derived from the model. The model is able to offer plausible accounts for common mental imagery phenomena and findings about the dual nature of imagery. The model implements an attention window to select regions of the visual representation for processing. The defined role of this structure can be used to predict eye movements during mental imagery tasks and as a novel way of evaluating theories and models of mental im-

agery.

Important aspects whose effects on the model's behavior needs to be investigated include working memory restrictions and similarly decay processes for both employed working memory structures. Furthermore, we are preparing appropriate eye tracking experiments to test the model's predictions about the construction sequence of multi-part images.

Acknowledgments

In this paper work done in the project R1-[ImageSpace] of the Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition is presented. Funding by the German Research Foundation (DFG) is gratefully acknowledged. We are thankful to the reviewers for their helpful comments.

References

Barkowsky, T. (2007). Modeling mental spatial knowledge processing: An AI perspective. In F. Mast & L. Jäncke (Eds.), *Spatial Processing in Navigation, Imagery, and Perception.* (p. 67-84). Berlin: Springer.

Bertel, S., Barkowsky, T., Engel, D., & Freksa, C. (2006). Computational modeling of reasoning with mental images: basic requirements. In D. Fum, F. del Missier, & A. Stocco (Eds.), *Proceedings of the 7th International Conference on Cognitive Modeling (ICCM 2006)* (p. 50-55). Edizioni Goliardiche; Trieste.

Chandrasekaran, B., Kurup, U., Banerjee, B., Josephson, J. R., & Winkler, R. (2004). An architecture for problem solving with diagrams. In A. Blackwell, K. Marriott, & A. Shimojima (Eds.), *Proceedings of Diagrams 2004* (pp. 151–165). Berlin: Springer.

Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1996). Object and Spatial Working Memory Activate Separate Neural Systems in Human Cortex. *Cereb. Cortex*, *6*(1), 39-49.

Farah, M. J., & Hammond, K. M. (1988). Visual and spatial mental imagery: Dissociable systems of representation. *Cognitive Psychology*, *20*, 439–462.

Finke, R. A. (1989). *Principles of mental imagery*. Cambridge, MA: MIT-Press.

Glasgow, J., & Papadias, D. (1992). Computational imagery. *Cognitive Science*, *16*, 355–394.

Gunzelmann, G., & Lyon, D. R. (2007). Mechanisms of human spatial competence. In T. Barkowsky, M. Knauff, G. Ligozat, & D. R. Montello (Eds.), *Spatial Cognition V - Reasoning, Action, Interaction* (p. 288-307). Springer Verlag; 14197 Berlin.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Johansson, R., Holsanova, J., & Holmqvist, K. (2005). What do eye movements reveal about mental imagery? Evidence from visual and verbal elicitations. In B. G. Bara, L. Barsalou, & M. M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (p. 1045 - 1059). Mahwah, NJ: Erlbaum.

Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.

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

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. New York: Oxford University Press.

Lathrop, S. (2008). *Extending cognitive architectures with spatial and visual imagery mechanisms*. Ph.d. thesis, University of Michigan.

Levine, D. N., Warach, J., & Farah, M. (1985). Two visual systems in mental imagery: Dissociation of "what" and "where" in imagery disorders due to bilateral posterior cerebral lesions. *Neurology*, *35*(7), 1010-.

Nestor, A., & Kokinov, B. (2004). Towards active vision in the DUAL cognitive architecture. *International Journal on Information Theories and Applications*, *11*, 9–15.

Schultheis, H., Barkowsky, T., & Bertel, S. (2006). LTM-C — An improved long-term memory for cognitive architectures. In D. Fum, F. del Missier, & A. Stocco (Eds.), *Proceedings of the 7th International Conference on Cognitive Modeling (ICCM 2006)* (p. 274 - 279). Edizioni Goliardiche; Trieste.

Sima, J. F., Lindner, M., Schultheis, H., & Barkowsky, T. (2010). Eye movements reflect reasoning with mental images but not mental models in orientation knowledge tasks. In C. Hölscher (Ed.), *Spatial Cognition VII* (pp. 248–261). Heidelberg: Springer Verlag.

Ungerleider, L., & Mishkin, M. (1982). Two cortical systems. *Analysis of Visual Behavior*, 549-586.