Itti, L., and Arbib, M.A., 2005, Attention and the Minimal Subscene, to appear in M. A. Arbib (Ed.), *Action to Language via the Mirror Neuron System*, Cambridge University Press.

Attention and the Minimal Subscene

Laurent Itti and Michael A. Arbib

ABSTRACT

We describe a computational framework that explores the interaction between focal visual attention, the recognition of objects and actions, and the related use of language. We introduce the notions of "minimal subscene" and “anchored subscene” to provide a middle ground representation, in which an agent is linked to objects or other agents via some action. We offer a preliminary model of visual attention which links bottom-up salience, contextual cues, object recognition, top-down attention, and short-term memory in building representations of subscenes. We then examine how this framework links to low-level visual perception, on the one end, and to sentences which describe a subscene or raise questions about the scene, on the other.

1. INTRODUCTION

The Mirror System Hypothesis (MSH), described in Chapter 1, asserts that recognition of manual actions may ground the evolution of the language-ready brain. More specifically, the hypothesis suggests that manual praxic actions provide the basis for the successive evolution of pantomime, then protosign and protospeech, and finally the articulatory actions (of hands, face and – most importantly for speech – voice) that define the phonology of language. But whereas a praxic action just *is* a praxic action, a communicative action (which is usually a compound of meaningless articulatory actions; see Goldstein et al., this volume, on duality of patterning) is *about something else*. We want to give an account of that relationship between the sign and the signified (Arbib, this volume, Section 4.3).

Words and sentences can be about many things and abstractions, or can have social import within a variety of speech acts. However, here we choose to focus our discussion by looking at two specific tasks of language in relation to a visually perceptible scene: (i) generating a description of the scene, and (ii) answering a question about the scene. At one level, vision appears to be highly parallel, whereas producing or understanding a sentence appears to be essentially serial. However, in each case there is both low-level parallel processing (across the spatial dimension in vision, across the frequency spectrum in audition) and high-level seriality in time (a sequence of visual

fixations or foci of attention in vision, a sequence of words in language). Our attempt to integrate these diverse processes requires us to weave together a large set of empirical data and computational models. The rest of this introduction is designed to provide a road map for the rest of the Chapter to help the reader integrate the varied strands of our argument.

Section 2: We introduce the notion of minimal and anchored subscenes as providing the link between the structure of a visual scene and the structure of a sentence which describes or queries it.

Section 3: We review the general idea of "cooperative computation" as well as of perceptual and motor schemas, and illustrate it with two classical models: the VISIONS model of recognizing a visual scene, and the HEARSAY model of understanding a spoken sentence.

Section 4: We also note two background studies of visual attention to which we have contributed, in both of which the notion of Winner Take All (WTA) plays a central role. The first, by Arbib & Didday, is more comprehensive in its conceptual structure; the second, by Itti & Koch, is restricted to low-level salience but has been worked out in detail and has yielded interesting results through detailed simulation.

Section 5: The unfolding of a description and the answering of a question require that attention be driven not solely by low-level salience but also by the search for objects (and actions) that are deemed relevant to the unfolding task. We thus present a conceptual model, based on a model implemented by Navalpakkam & Itti, which builds on the concepts of Section 4 but with the addition of a Task Relevance Map (TRM), Short-Term Memory (STM), top-down task-dependent biasing, and other subsystems. We make explicit the role of minimal and anchored subscenes in this processing, in particular by showing how continuous scene analysis yields sequences of temporally less volatile short-term memory representations that are amenable to verbalization.

Section 6: But of course we are not the first to investigate the relevance of visual attention to language. Henderson & Ferreira offer a rich collection of relevant review articles. We briefly note the contributions of Tanenhaus, Griffin and Bock.

Section 7: The theoretical framework which comes closest to what we seek is that offered by Knott, which builds on the Itti-Koch model and links this to a system for execution and observation of actions. Knott translates the sensorimotor sequence of attention to the scene into the operations involved in constructing the syntactic tree for its description. We argue that the theory is more pertinent if we build the scene description on symbolic short-term memory, with items tagged for relevance, rather than on the eye movements that went into the building of that representation. Competitive queuing yields the sequence of "virtual attention shifts" for this internal representation.

Section 8: How can some of the new notions introduced in this chapter be explored experimentally? One difficulty is that the framework is oriented more towards dynamic scenes, in which change constantly occurs, than towards static images which may more easily be studied experimentally. To illustrate how our framework may prompt for new experiments on human comprehension of dynamic scenes, we conducted a pilot study of a person describing a range of videoclips and analyzed the relation between the subject's eye movements and the descriptions he generated.

In the remainder of the chapter, we review other research efforts relevant to developing a comprehensive computational model adequate to meeting the challenges of the data and modeling introduced above.

Section 9: We note interesting efforts (VITRA, Nevatia et al.) within AI on the recognition of events in dynamically changing scenes.

Section 10: We link Knott's brief description of motor control to the more general notion of forward and inverse models, then briefly summarize the discussion given by Oztop et al. (this volume) relating the FARS and MNS models of manual action and action recognition, respectively, to this general framework. We return to the distinction between the sign and the signified to distinguish between producing or perceiving a word and perceiving the concept that it signifies.

Section 11: We also note the preliminary work by Vergnaud and Arbib giving a bipartite analysis of the verb in terms of mirror neurons and canonical neurons.

Section 12: Both Knott and Vergnaud operate within the framework of generative grammar. We review the attractions of an alternative framework, construction grammar (employed by Kemmerer, this volume), for our work. We show how "vision constructions" may synergize with "grammar constructions" in structuring the analysis of a scene in relation to the demands of scene description and question answering in a way which ties naturally into our concern with minimal and anchored subscenes.

Section 13: We revisit SVSS to discuss its extension to extract episodes from dynamically changing visual input. While the detailed discussion of language processing is outside the scope of this chapter, we do suggest how the integration of ideas from Levelt's model of language production and the HEARSAY model of speech understanding may set the stage for a cooperative computation model of language perception and production. We then outline how this might be integrated with our general analysis of attention-based visual perception, and of the generation and recognition of actions, in defining a truly general model for linking eye movements to the processes of scene description and question answering.

Section 14: We recall a highly schematic diagram developed by Arbib & Bota (2003) which sketches the relations between action generation, action recognition and language within the anatomical framework offered by the Mirror System Hypothesis. Where this diagram focuses on single actions, we review the insights offered by the functional analysis in this chapter towards extending the Mirror System Hypothesis to complex scenes.

2. MINIMAL AND ANCHORED SUBSCENES

In this chapter, we outline a computational framework in which to probe the visual mechanisms required to recognize salient aspects of the environment, whether as a basis for praxic action (e.g., manipulation of objects or locomotion), or as part of the linguistic acts of scene description or question answering. The key concept introduced for this purpose is that of a *minimal subscene* as the basic unit of action recognition, in which an agent interacts with objects or others. We now define minimal subscenes and their expansion to anchored subscenes, and then briefly review processes of active, goal-oriented scene perception to be treated at greater length in the rest of the chapter.

2.1. Minimal subscenes defined

While the definition of what constitutes a minimal subscene is not fully rigorous, the notion is that, for whatever reason, an agent, action, or object may attract the viewer's attention more strongly than other elements which may be present in the scene. If that agent, action, or object is of interest by some measure, then attention will be directed "top-down" to seek to place this "anchor" in context within the scene:

- Given an *agent* as "anchor", complete the minimal subscene by including the actions in which the agent is engaged and the objects and (where appropriate) other agents engaged in these actions;
- Given an *action* as "anchor", complete the minimal subscene by including the agents and objects involved in that action;
- Given an *object* as "anchor", complete the minimal subscene by including the actions performed on that object and the agents and (where appropriate) other objects engaged in that action.

Importantly, which anchor is of interest and which minimal subscene is incrementally constructed during observation is dependent upon tasks and goals, so that the same visual input may yield different minimal subscenes depending on these key top-down factors. The minimal subscene thus is fundamentally an observer-dependent notion. Data observations only enrich it to the extent that they are relevant to the tasks and goals of the observer. Also noteworthy is that our definition does not require that all the elements of the minimal subscene be present in the visual world at every instant; thus, it essentially is a *short-term memory* construct, which evolves at a somewhat slower pace than the constantly streaming visual inputs that nourish it.

A minimal subscene as defined above may be extended by adding details to elements already in the subscene, or by adding new elements which bear some strong relationship to elements in the subscene. We refer to the result as an *anchored subscene* since it is defined by expanding the relationships which link elements of the original subscene to its anchor. However, attention may be caught by a salient object or action that is not part of the original subscene and that is not related or relevant to the current subscene, yet triggers the interest of the observer. This may anchor a new subscene to be added to the observer's inherently dynamic short term memory (STM). Thus, by definition, minimal or anchored subscenes must at least include some neural representations for or links to *the collection of featural representations (appearance, identity, properties), spatial representations (location, size, relationships) and dynamic representations (actions linking actors, objects, and actions) that have been observed and attended to in the recent past, and that have further been retained as potentially relevant to the current tasks and goals of the observer*. This new subscene may either coexist with the previous one, for example if the actors of the previous subscene remain present, or it may take precedence and become the only current minimal subscene, for example after the elements of the previous subscene have drifted out of the field of view. Internal representations of the scene thus develop in STM and are refined as new data are accumulated by the sensors and evaluated against the internal beliefs and goals of the observer. Conversely, aspects of STM may be discarded either due to some process of forgetting, or because they are no longer relevant to the goals of the system or its changing relationship with the environment.

A significant contribution of the present work is to examine how these notions, which we have thus far introduced in relation to perception, link to sentences which describe a subscene or raise questions about the scene. We view the minimal subscene as providing a middle-ground representation between patterns of neural activation in response to the observation of a dynamic visual scene involving some actors and actions, and more abstract symbolic descriptions of the scene in some human language. We study how such an *action-object frame* relates to the *verb-argument structure* for "Who is doing what and to whom". For example, if Harry's forearm moves up and down as he holds a hammer, then we may have an action Hit(Harry, Nail, Hammer) which falls under the general action-object frame of Action(Actor, Patient, Instrument) and is expressible in English as "Harry is hitting the nail with a hammer." Of course, there are other ways to express this same idea in English, such as "Harry hammered the nail." As made clear by our tripartite account of "minimal subscene", the viewer's attention might first be focused on Harry, say, and extend to recognition of his action and the objects involved, or start with the hammering

movement and extend this to complete the minimal subscene. The anchor of the scene will in general determine the focus of the sentence.

2.2. Processes of active, goal-oriented scene perception

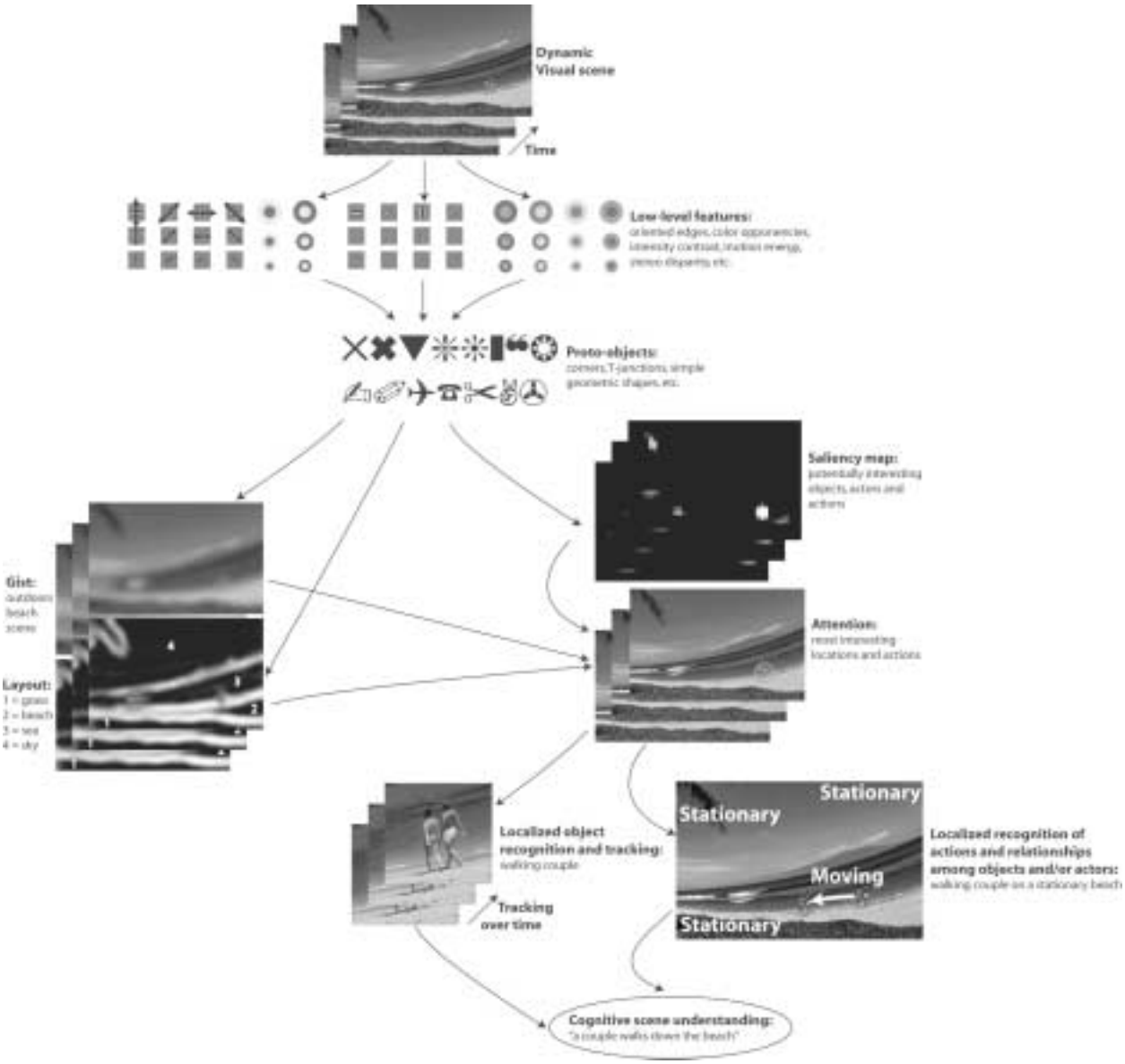


Figure 1: Overview of how task influences visual attention: It primes the desired features that are in turn used to compute the gist, layout, and the bottom-up salience of scene locations. Further, task influences top-down processes that predict the relevance of scene locations based on some prior knowledge. Finally, the gist, layout, bottom-up salience and top-down relevance of scene locations are somehow combined to decide the focus of attention.

Our understanding of the brain mechanisms at play during active, goal-oriented scene perception has significantly progressed over recent years, through psychophysical, electrophysiological, imaging and modeling studies. The current state of understanding is sketched as in Figure 1. Humans rapidly extract a wealth of

information, often referred to as the "gist of the scene", from the first glance at a new scene. We use the term gist in the sense of some overall classification – such as "a beach scene", "a landscape", "a suburban street scene", or "a battle zone" – that can provide spatial priors as to the most likely regions of interest. Gist thus refers to scene representations which are acquired over very short time frames, for example as observers view briefly flashed photographs and are asked to describe what they see (Potter, 1975; Biederman, 1982; Oliva & Schyns, 1997).[1] With very brief exposures (100ms or below), gist representations are typically limited to a few general semantic attributes (e.g., indoors, outdoors, office, kitchen) and a coarse evaluation of distributions of visual features (e.g., highly colorful, grayscale, several large masses, many small objects) (Sanocki & Epstein, 1997; Rensink, 2000). Gist may be computed in brain areas which have been shown to preferentially respond to types of "places," that is, visual scene types with a restricted spatial layout (Epstein & Kanwisher, 2000). Spectral contents and color diagnosticity have been shown to influence gist perception (Oliva & Schyns, 1997; 2000), leading to the development of computational models that approximate the computation of gist by classifying an image along a number of semantic and featural dimensions, based on a one-shot global analysis of the spectral composition of a scene (Torralba, 2003).

In our conceptual framework, the gist representation cooperates with visual attention to focus on cognitively salient visual targets. The result will be updating of the minimal or anchored subscene representation after each shift of attention to a new scene element that is related to elements already represented within that subscene, or the initiation or updating of another subscene in case no such relation exists. The focusing of attention onto specific scene elements has been shown to be mediated by the interplay between bottom-up (dependent upon uninterpreted properties of the input images) and top-down factors (which bring volition, expectations, and internal beliefs of the observers into play). Attention is thus attracted bottom-up towards conspicuous or salient scene elements (e.g., a bright flickering light in an otherwise dark environment) and top-down towards locations which the observer believes are or soon will be important (e.g., the expected endpoint of a ball's trajectory). We posit that this interplay between reactive and proactive modes of attention plays a critical role in the elaboration of the minimal subscene. Indeed, studies of change blindness (Rensink et al. 1997) and inattentional blindness (Simons, 2000) suggest that human observers are unlikely to notice or report scene elements which an outsider might expect to gain their attention, either because they were distracted by salient bottom-up stimuli, or because they were strongly focused

[1] This first-impression notion of gist is to be contrasted with the sense of gist as summarizing the key aspects of a scene that can only be extracted after careful analysis, as in "John is finally proposing to Mary but she doesn't seem happy about it".

top-down onto a visual inspection task. Thus, attention may gate the addition of new constituents to subscene representations.

Here we must add that, while much of present visual scene analysis focuses on the recognition of elements in a single static view, our real concern (as becomes clear in Sections 8 and 9) is with dynamically changing scenes. In such a framework, the gist will play a broader role being updated as the dynamics of the scene and its interpretation unfold.

Recognizing objects, actors and actions is another obvious necessary step in the construction of subscenes. Unless an attended entity is recognized, it will be difficult to evaluate how relevant it is to current behavioral goals, and whether it should be immediately discarded or integrated into the current subscene representation. Beyond the instantaneous recognition of the scene element that is the current focus of attention, some memory trace of that element, if deemed relevant and recognized with high confidence, must be stored if the subscene is to be built incrementally by integrating across attention and gaze shifts, and possibly later reported on. Many studies have explored the mechanisms and limitations of the short-term memory processes involved in transiently holding some representation of the constituents of the minimal subscene. These studies have yielded a wide range of putative representations and levels of detail, ranging from the "world as an outside memory" hypothesis (O'Regan, 1992) where virtually no visual details are retained internally in memory because they could be easily fetched back from the outside world itself if desired, to collections of 4-6 "object files" (Treisman & Gelade, 1980; Irwin & Andrews, 1996; Irwin & Zelinsky, 2002) which crudely abstract the visual details of retained scene elements, to more complete representations with rich detail (Hollingworth, 2004; Hollingworth & Henderson, 2002).

Finally, cognition and reasoning must influence the construction of the minimal subscene in at least two ways. First, they determine in real-time the evolution of top-down influences on attention, by evaluating previous objects of attention and recognition against goals, and inferring a course of action. For example, if your attention was just caught by a man's face, but you want to know about his shoes, just look down. Second, in many situations, it seems intuitively reasonable to assume that not all attended and recognized scene elements will automatically become part of the minimal subscene. Instead, some cognitive evaluation of the behavioral relevance of an object of attention is likely to play a role in deciding whether it will be retained or immediately forgotten. For example, as you scan a crowd for a loved one, you will probably forget about most of the persons who you attended to but who were false positives during your search. Many factors, possibly including familiarity (e.g., what is my cousin doing in that crowd?) or an element of surprise (who is that man with a green face?) may augment evaluation of attended scene

elements when determining whether or not they will be integrated into the minimal subscene representation or barred from it and forgotten.

This brief overview suggests a highly active and dynamic view of goal-oriented scene perception. Instead of a snapshot of the scene such as gist may be, the minimal subscene – and the anchored subscene which extends it – is an evolving representation, shaped by a combination of attending, recognizing, and reasoning over time. Before diving into the details of how the orchestration of all the factors just mentioned may contribute to the minimal subscene, one challenge is to define a suitable neuro-cognitive framework to support its implementation.

3. COOPERATIVE COMPUTATION AND SCHEMA THEORY

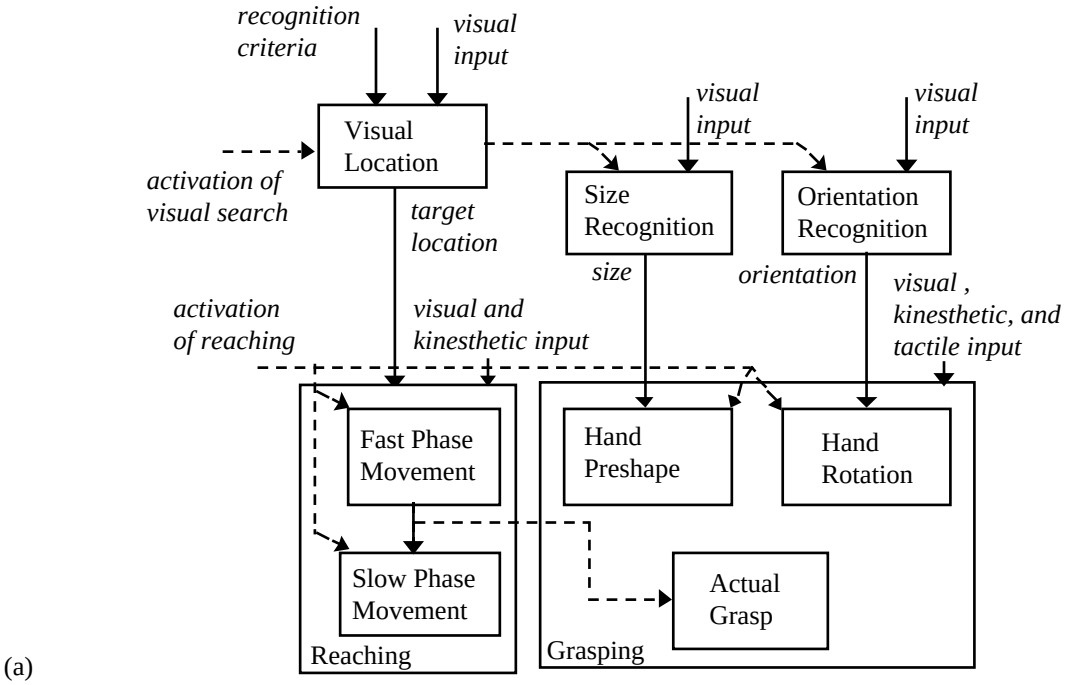
Visual organization in the primate brain relies heavily on topographic neural maps representing various transformed versions of the visual input, distributed population coding for visual features, patterns, or objects present in the scene, and highly dynamic representations shaped both bottom-up (by visual inputs) and top-down (by internal cognitive representations). To relate these highly redundant, distributed and dynamic neural representations to corresponding linguistic representations, we appeal to the general idea of "cooperative computation" as well as the perceptual and motor schemas of "schema theory", and illustrate them with two classical models: the VISIONS model of recognizing a visual scene, and the HEARSAY model of understanding a spoken sentence.

3.1. Schemas for Perceptual Structures and Distributed Motor Control

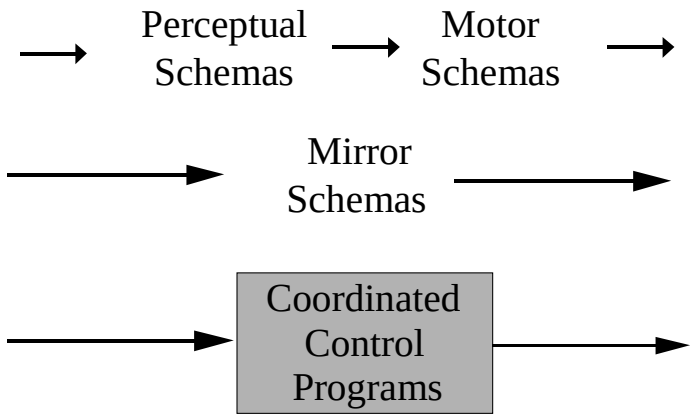
Arbib (1981; Arbib et al., 1998, Chapter 3) offers a version of *schema theory* to complement neuroscience's terminology for levels of structural analysis with a framework for analysis of function in terms of *schemas* (units of functional analysis). Central to our approach is *action-oriented perception*, as the active "organism" (which may be an animal or an embodied computational system) seeks from the world the information it needs to pursue its chosen course of action. A *perceptual schema* not only determines whether a given "domain of interaction" (an action-oriented generalization of the notion of object) is present in the environment but can also provide parameters concerning the current relationship of the organism with that domain. *Motor schemas* provide the control systems which can exploit such parameters and can be coordinated to effect the wide variety of action.

A schema is, in basic cases, what is learned about some aspect of the world, combining knowledge with the processes for applying it; a brain may deploy more than one *instance* of the processes that define a given schema. In particular, schema instances may be combined (possibly including those of more abstract schemas, including coordinating schemas) to form *schema assemblages*. For example, an assemblage of perceptual schema instances may combine an estimate of environmental state with a representation of goals and needs. A *coordinated control*

program is a schema assemblage which processes input via perceptual schemas and delivers its output via motor schemas, interweaving the activations of these schemas in accordance with the current task and sensory environment to mediate more complex behaviors.



(a)



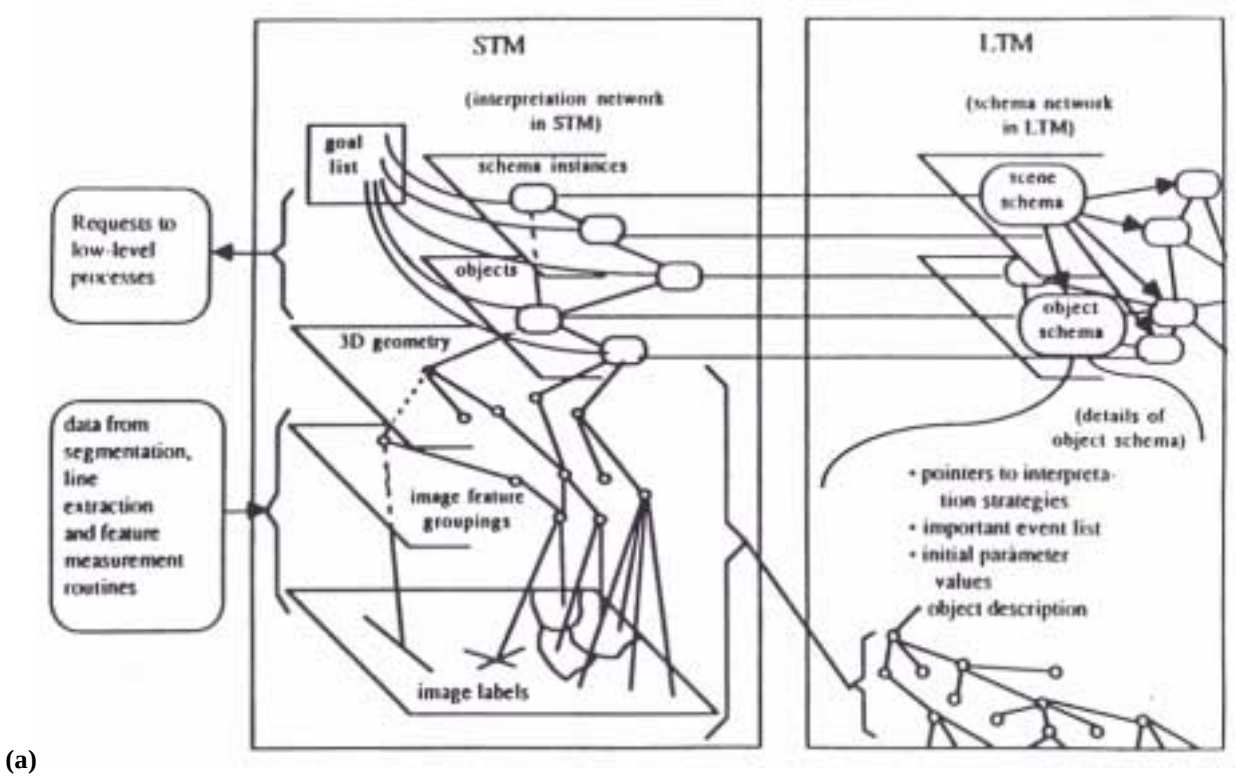
(b)

Figure 2: (a) Hypothetical coordinated control program for reaching and grasping. Different perceptual schemas (top half of figure) provide input for the motor schemas (bottom half of figure) for the control of "reaching" (arm transport ≈ reaching) and "grasping" (controlling the hand to conform to the object). Note too the timing relations posited here between subschemas within the " Reaching" motor schema and those within the motor schema for "Grasping". Dashed lines - activation signals; solid lines - transfer of data. (Adapted from Arbib 1981.) (b) A general diagram emphasizing that the linkage between perception and action may involve perceptual schemas talking to motor schemas, a “mirror” schemas which can both recognize and generate a class of actions (integrating a perceptual and motor schema in one), or coordinated control programs which are assemblages of perceptual, motor and coordinating schemas.

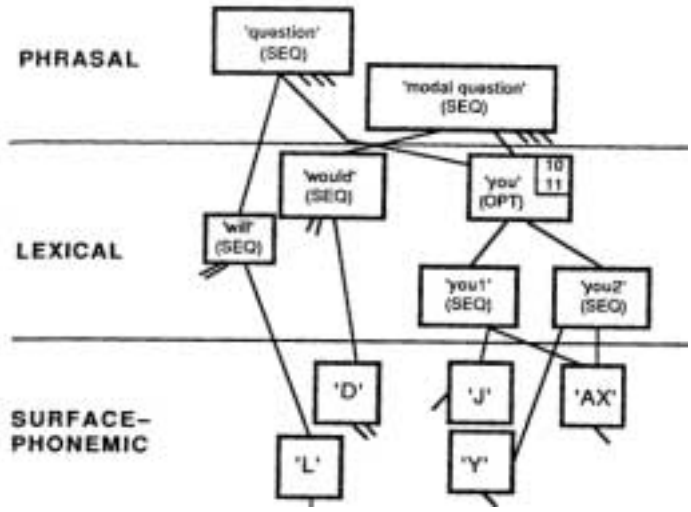
Figure 2(a) shows the original coordinated control program (Arbib 1981, analyzing data of Jeannerod & Biguer 1982). As the hand moves to grasp an object, it is *preshaped* so that when it has almost reached the object, it is of the right shape and orientation to enclose some part of the object prior to gripping it firmly. Moreover (to a first approximation; see, e.g., Hoff & Arbib, 1993, for a revised analysis), the movement can be broken into a fast phase and a slow phase. The output of three perceptual schemas is available for the control of the hand movement by concurrent activation of two motor schemas: *Reaching* controls the arm to transport the hand towards the object and *Grasping* first preshapes the hand. Once the hand is preshaped, it is (according to Figure 2) only the completion of the fast phase of hand transport that "wakes up" the final stage of *Grasping* to shape the fingers under control of tactile feedback. (This model anticipates the much later discovery of perceptual schemas for grasping in a localized area [AIP] of parietal cortex and motor schemas for grasping in a localized area [F5] of premotor cortex. See Chapter 1 and below.) Jeannerod (1997) surveys the role of schemas and other constructs in the cognitive neuroscience of action; schemas have also played an important role in the development of behavior-based robots (Arkin, 1998).

Figure 2(a) clearly separates perceptual and motor schemas. But this raises the question as to why I do not combine perceptual and motor schemas into a single notion of schema that integrates sensory analysis with motor control. Indeed, there are cases where such a combination makes sense. However, recognizing an object (an apple, say) may be linked to many different courses of action (to place it in one's shopping basket; to place it in a bowl; to pick it up; to peel it; to cook with it; to eat it; to discard a rotten apple, etc.). Of course, once one has decided on a particular course of action then specific perceptual and motor subschemas must be invoked. But note that, in the list just given, some items are apple-specific whereas others invoke generic schemas for reaching and grasping. It was considerations like this that led me Arbib (1981) to separate perceptual and motor schemas – a given action may be invoked in a wide variety of circumstances; a given perception may precede many courses of action. There is no one grand "apple schema" which links all "apple perception strategies" to "every action that involves an apple". Moreover, in the schema-theoretic approach, "apple perception" is not mere categorization – "this is an apple" – but may provide access to a range of parameters relevant to interaction with the apple at hand. Thus this approach views the brain as encoding a varied network of perceptual and motor schemas and coordinated control programs built upon them, perhaps with the mediation of coordinating schemas. Only rarely (as in the case of certain basic actions) will the perceptual and motor schemas be integrated into a "mirror schema" (Figure 2b).

3.2. VISIONS: Schemas for Visual Scene Understanding



(a)



(b)

Figure 3. (a) The VISIONS paradigm for cooperative computation in visual scene analysis. Interpretation strategies are stored in schemas which are linked in a schema network in Long Term Memory (LTM). Under the guidance of these schemas, the intermediate representation (data concerning edges, region boundaries, color, texture, etc.) is modified and interpreted by a network of schema instances which label regions of the image and link them to a 3D geometry in Short Term Memory (STM). [From Arbib 1989, after Weymouth, 1986.] (b) The HEARSAY paradigm for cooperative computation in speech understanding.

An early example of schema-based interpretation for visual scene analysis in the VISIONS system (Draper et al., 1989; Arbib, 1989, Sec. 5.3). While this is an "old" system, it allows us to describe how distributed computation may be used in visual scene perception in a way that we can build upon in presenting our current approach to minimal subscenes and language. In VISIONS, there is no extraction of gist – rather, the gist is prespecified so that only those schemas are deployed relevant to recognizing a certain kind of scene (e.g., an outdoor scene with houses, trees, lawn, etc.). Low-level processes take an image of such an outdoor visual scene and extract and builds a representation in the *intermediate database* – including contours and surfaces tagged with features such as color, texture, shape, size and location. An important point is that the segmentation of the scene in the intermediate database is based not only on bottom-up input (data-driven) but also on top-down hypotheses (e.g., that a large region may correspond to two objects, and thus should be resegmented; or that two continuous regions may correspond to parts of the same object and should be merged). These are the features on which bottom-up attention (see Figure 5 below) can operate, but VISIONS has a limited stock of schemas and so applies perceptual schemas across the whole intermediate representation to form confidence values for the presence of objects like houses, walls and trees. The knowledge required for interpretation is stored in LTM (long-term memory) as a network of schemas, while the state of interpretation of the particular scene unfolds in STM (short-term or working memory) as a network of schema instances (Figure 3a). Note that this STM is not defined in terms of recency but rather in terms of continuing relevance. Our interest in VISIONS is that it provides a good conceptual basis for the elaboration of the concept of minimal subscene, although missing crucial components that include attention, representations of dynamic events and relationships, and rapidly changing top-down modulation as an increasing number of attended scene elements are evaluated against goals and tasks.

In the VISIONS system, interpretation of a novel scene starts with the bottom-up instantiation of several schemas (e.g., a certain range of color and texture might cue an instance of the foliage schema for a certain region of the image). When a schema instance is activated, VISIONS links it with an associated area of the image and an associated set of local variables. Each schema instance in STM has an associated confidence level which changes on the basis of interactions with other units in STM. The STM network makes context explicit: each object represents a context for further processing. Thus, once several schema instances are active, they may instantiate others in a "top-down" way (e.g., recognizing what appears to be a roof will activate an instance of the house schema which will in turn activate an instance of the wall schema to seek confirming evidence in the region below that of the putative roof). Ensuing computation is based on the competition and cooperation of concurrently active schema instances.

Once a number of schema instances have been activated, the schema network is invoked to formulate hypotheses, set goals, and then iterate the process of adjusting the activity level of schemas linked to the image until a coherent scene interpretation of (part of) the scene is obtained. Cooperation yields a pattern of "strengthened alliances" between mutually consistent schema instances that allows them to achieve high activity levels to constitute the overall solution of a problem. As a result of competition, instances which do not meet the evolving consensus lose activity, and thus are not part of this solution (though their continuing subthreshold activity may well affect later behavior). Successful instances of perceptual schemas become part of the current short-term model of the environment.

The classic VISIONS system had only a small number of schemas at its disposal, and so could afford to be lax about scheduling their application. However, for visual systems operating in a complex world, many schemas are potentially applicable, and many features of the environment are interpretable. In this case, “attention” – the scheduling of resources to process specific parts of the image in particular ways – becomes crucial. We emphasize that attention includes not only *where* to look but also *how* to look. We shall return to the theme of attention in Section 5, but first let us place this notion of cooperative computation in a broader perspective.

Returning to Figure 3a, we make three points:

a) We regard Gist as priming the appropriate “top-level” schema – e.g., suburban scene, city scene, beach scene, etc.

b) We note that the lower levels (“image feature groupings” – basically, the “scene layout”) are very close to the “intermediate database” – with local features replaced by interpretable regions which can be linked (with more or less confidence) to the (parameterized) schema instances that constitute their interpretation. 3D geometry may emerge from this, whether linked to a schema instance (the shape of a house) or not (as in the unfolding of landscape).

c) We note the need to interpose a level below object schemas for “prototype objects” – feature patterns that are relatively easy to detect bottom-up, yet which greatly focus the search for schemas compatible with a given region (cf. Rensink, 2000).

d) We stress that STM does not hold a unique schema instance for each region. Rather, schema instances may compete and cooperate with shifting confidence values till finally a group of them passes some threshold level of confidence to constitute the interpretation of the scene.

e) Finally, we will below distinguish between two aspects of short-term memory (STM): the first is very close to the intermediate database used by VISIONS, which evolves rather rapidly, fairly closely following any changes in the incoming visual inputs; the other is a more symbolic component, which is the basis for minimal and anchored subscene representations, evolving at a slower pace as either new scene elements both confidently identified and deemed relevant to the current task enrich the subscene, or elements previously in the subscene but no longer relevant fade away.

3.3. From Vision to Action

We now "reflect" VISIONS, analyzing the scene in terms of opportunities for action – motor schemas which then compete for realization (Arbib and Liaw, 1995). In addition, motor schemas are affected "top down" by goals and drive states, and "middle out" by the priming effect of other motor schemas. While only a few perceptual schemas may be active for the current focus of attention, STM will be updated as new results come in from this focal processing. STM is now more dynamic and task-oriented and must include a representation of goals and needs, linking instances of perceptual schemas to motor schemas, providing parameters and changing confidence levels, so as to provide suitable input to STM. As their activity levels reach threshold, certain motor schemas create patterns of overt behavior. To see this, consider a driver instructed to "Turn right at the red barn". At first the person drives along looking for something large and red, after which the perceptual schema for barns is brought to bear. Once a barn is identified, the emphasis shifts to recognition of spatial relations appropriate to executing a right turn "at" the barn, but constrained by the placement of the roadway, etc. The latter are an example of *affordances* in the sense of Gibson (1979), i.e., information extracted from sensory systems concerning the possibility of interaction with the world, as distinct from recognition of the type of objects being observed. For example, optic flow may alert one to the possibility of a collision without any analysis of what it is that is on a collision course.

In the VISIONS system, schemas in LTM are the passive codes for processes (the programs for deciding if a region is a roof, for example), while the schema instances in STM are active copies of these processes (the execution of that program to test a particular region for "roofness"). By contrast, it may be that in analyzing the brain, we should reverse the view of activity/passivity of schemas and instances: the active circuitry is the *schema*, so that only one or a few instances can apply data-driven updating at a time, while the *schema instance* is a parameterized working memory of the linkage of a schema to a region of the scene, rather than an active process.

Such considerations offer a different perspective on the neuropsychological view of working memory offered by Baddeley (2003). The initial three-component model of working memory proposed by Baddeley and Hitch (1974)

posits a *central executive* (an attentional controller) coordinating two subsidiary systems, the *phonological loop*, capable of holding speech-based information, and the *visuospatial sketchpad*. The latter is viewed as passive, since the emphasis of Baddeley's work has been on the role of working memory in sentence processing. Baddeley (2003) added an episodic LTM to the Baddeley-Hitch model, with the ability to hold language information complementing the phonological loop and (the idea is less well developed) an LTM for visual semantics complementing the visuospatial sketchpad. He further adds an *episodic buffer*, controlled by the central executive, which is assumed to provide a temporary interface between the phonological loop and the visuospatial sketchpad and LTM. The Arbib-Liaw scheme seems far more general, because it integrates dynamic visual analysis with the ongoing control of action. As such, it seems better suited to encompass Emmorey's notion (see the Section "Broca's Area and Working Memory for Sign Language" in Emmorey 2004) that sign language employs a visuospatial phonological short-term store. With this, let us see how the above ideas play out in the domain of speech understanding.

3.4. HEARSAY: Schemas for Speech Understanding

Jackendoff (2002) makes much use of the AI notion of blackboard in presenting his architecture for language. HEARSAY-II (Lesser et al., 1975) was perhaps the first AI system to develop a blackboard architecture, and the architecture of the VISIONS computer vision system was based on the HEARSAY architecture as well as on neurally-inspired schema theory. While obviously not the state of the art in computer-based speech understanding, it is of interest here because it foreshadows features of Jackendoff's architecture. Digitized speech data provide input at the *parameter level*; the output at the *phrasal level* interprets the speech signal as a sequence of words with associated syntactic and semantic structure. Because of ambiguities in the spoken input, a variety of hypotheses must be considered. To keep track of all these hypotheses, HEARSAY uses a dynamic global data structure, called the *blackboard*, partitioned into various levels; processes called *knowledge sources* act upon hypotheses at one level to generate hypotheses at another (Figure 3b). First, a knowledge source takes data from the *parameter level* to hypothesize a phoneme at the *surface-phonemic level*. Many different phonemes may be posted as possible interpretations of the same speech segment. A lexical knowledge source takes phoneme hypotheses and finds words in its dictionary that are consistent with the phoneme data – thus posting hypotheses at the *lexical level* and allowing certain phoneme hypotheses to be discarded. These hypotheses are akin to the schema instances of the VISIONS system (Figure 3a).

To obtain hypotheses at the *phrasal level*, knowledge sources embodying syntax and semantics are brought to bear. Each hypothesis is annotated with a number expressing the current confidence level assigned to it. Each

hypothesis is explicitly linked to those it supports at another level. Knowledge sources cooperate and compete to limit ambiguities. In addition to data-driven processing which works upward, HEARSAY also uses hypothesis-driven processing so that when a hypothesis is formed on the basis of partial data, a search may be initiated to find supporting data at lower levels. For example, finding a verb that is marked for plural, a knowledge process might check for a hitherto unremarked "s" at the end of a preceding noun. A hypothesis activated with sufficient confidence will provide context for determination of other hypotheses. However, such an *island of reliability* need not survive into the final interpretation of the sentence. All we can ask is that it forwards the process which eventually yields this interpretation.

Arbib and Caplan (1979) discussed how the knowledge sources of HEARSAY, which were scheduled serially, might be replaced by schemas distributed across the brain to capture the spirit of "distributed localization" of Luria (e.g., 1973). Today, advances in the understanding of distributed computation and the flood of brain imaging data make the time ripe for a new push at a neurolinguistics informed by the understanding of cooperative computation. It is also worth relating the discussion of VISIONS and HEARSAY to the evolutionary theme, "from action to language", of the present volume. While non-humans have well-developed mechanisms for scene analysis and for integrating that analysis to their ongoing behavior, they lack the ability to link that capability for "praxic behavior" to "communicative behavior" that can convert the perception of agents, actions, and objects and their relationships with each other and ongoing behavior into a structured set of symbols which can be used to express certain details of these relationships to others. Thus while Figure 3 illustrates the parallels between the blackboard architectures for visual (VISIONS) and speech (HEARSAY) perception, the evolution of human brain mechanisms that support language perception and production by exploiting variants of the mechanisms of perception and production relevant to praxic action is in no sense direct.

We close this Section by summarizing its main contributions. Section 3.1 introduced the general framework of schema theory (perceptual and motor schemas; schema assemblages; coordinated control programs). Section 3.2 showed how the VISIONS system could represent a static visual scene by a hierarchically structured network of schema instances linked to regions of the image. In Section 3.3, we suggested how the framework offered by VISIONS might be extended to include motor schemas and the planning of action, as well as perceptual schemas for scene recognition. Complementing VISIONS, the HEARSAY system (Section 3.4) demonstrated how to represent a speech stream by a hierarchically structured network of schema instances linked to time intervals of the spoken input. The role of attention in scheduling resources for scene recognition was implicit in our description of how in

VISIONS activation of a roof schema instance might lead to activation of an instance of the wall schema to check the region below that linked to the roof schema. In the next section, we present two "classic" models which make explicit the role of attention in vision. We then build on this in Figure 5 to present a framework for modeling the goal-directed and action-oriented guidance of attention. Here the emphasis is on recognition of a single static scene. Our challenge in later sections will be to combine the insights of these two classical systems to described how to represent a dynamic visual scene (or multimodal sensory data more generally) by a hierarchically structured network of schema instances each linked to a space-time region of the image. For example, a person may only be tracked over a certain time interval as they move from place to place; an action will extend across a certain region of time but may be quite localized.

4. INTEGRATING ATTENTION AND DYNAMIC SCENE ANALYSIS

We now turn to two background studies of visual attention to which we have contributed, in both of which the notion of Winner Take All (WTA) plays a central role. The first, by Arbib & Didday, is more comprehensive in its conceptual structure; the second, by Itti & Koch, is restricted to low-level salience but has been worked out in detail and has yielded interesting results through detailed simulation.

4.1. The Arbib-Didday Model

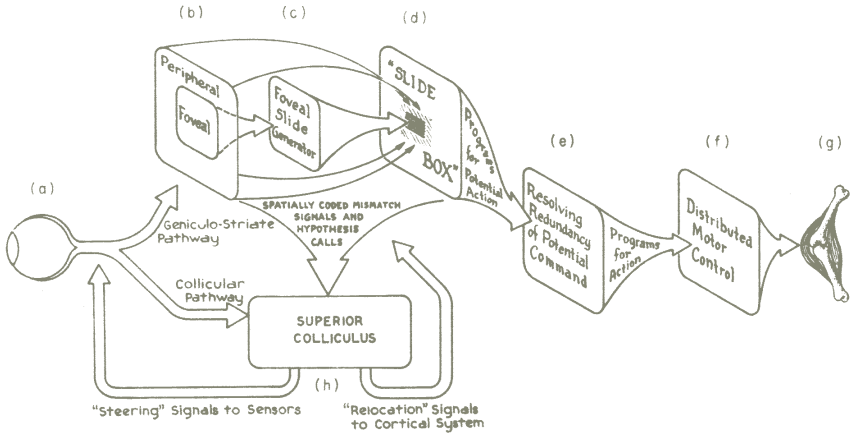


Figure 4. A "two visual system" model of visual perception (Arbib and Didday, 1975)

Our ancient "two visual system" model of visual perception (Arbib and Didday, 1975) has the following components:

1) The superior colliculus (h) responds to visual input to direct the eyes (a) to turn to foveate a new target. The target is chosen by a winner-take-all mechanism (inspired by frog prey selection) in response to bottom-up

salience cues (as developed in more detail in the Itti-Koch model of the next section) and top-down attention cues (as developed in the SVSS model of Section 5, below).

2) Arbib and Didday use the terms "slide" and "slide box" (d) where we would now speak of "schemas" and "STM" (a "linked collection of relevant schemas", as in VISIONS). In fact, the slide-box here may be seen as combining the functions of the intermediate database and STM in VISIONS. Foveal attention (b) provides the detailed input to appropriately add slides to, or adjust slides (c) in, the slide box.

3) As the eyes move, the foveal input must be redirected appropriately [(c) → (d)] to address the appropriate spatially tagged region of the slide box (cf. the notion of "dynamic remapping"; Dominey & Arbib, 1992; Medendorp et al., 2003). Moreover, the model of Figure 4 is designed to process dynamic scenes, so that the state of the slide-box will depend as much on the retention of earlier hypotheses as on the analysis of the current visual input.

4) A crucial point (anticipating Arbib and Liaw, 1995) is that perceptual schemas are not ends in themselves but are linked to motor schemas for potential action (the other half of (d)) with respect to entities in the observed scene.

5) Since more motor schemas may initially be activated than can be executed, a process of "resolving redundancy of potential command" (McCulloch, 1965) is required (e) to determine which actions are indeed executed (f, g) – just as the superior colliculus (h), the "motor side" of vision in the present model, must execute a winner-take-all computation to select the next focus of attention.

4.2. The Itti-Koch Model

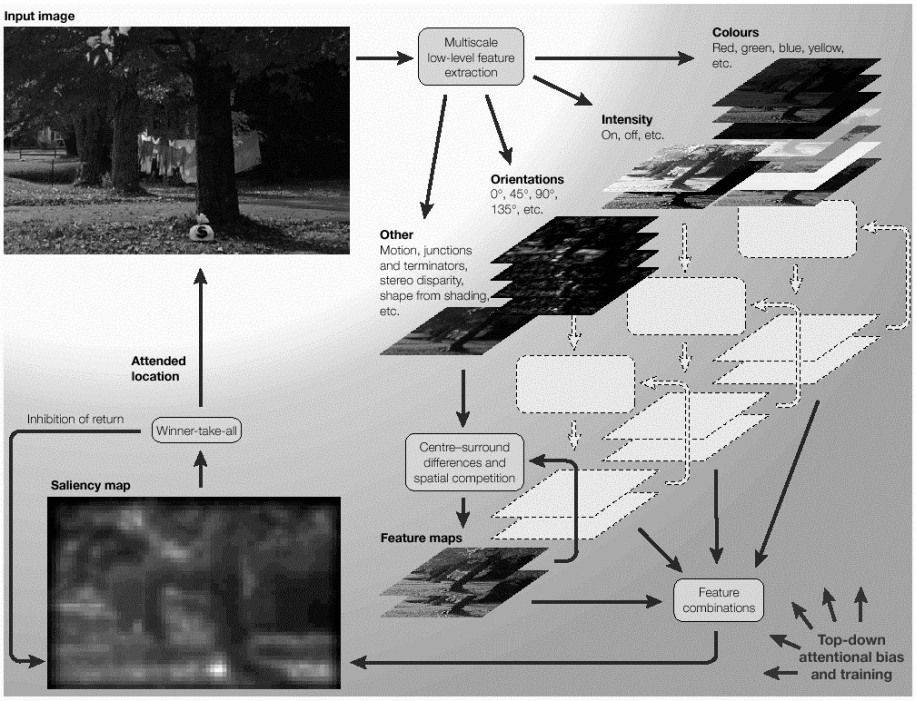


Figure 5: Overview of Itti & Koch's (2001, 2001) model of visual attention guided by bottom-up salience.

In our neuromorphic model of the bottom-up guidance of attention in primates (Figure 5; Itti & Koch, 2000, 2001), the input video stream is decomposed into eight feature channels at six spatial scales. After surround suppression, only a sparse number of locations remain active in each map, and all maps are combined into the unique *saliency map*. This map is scanned by the focus of attention in order of decreasing saliency through the interaction between a winner-take-all mechanism (which selects the most salient location) and an inhibition-of-return mechanism (which transiently suppresses recently attended locations from the saliency map). Because it includes a detailed low-level vision front-end, the model has been applied not only to laboratory stimuli, but also to a wide variety of natural scenes (e.g., Itti *et al.*, 2001), predicting a wealth of data from psychophysical experiments.

In this model, inhibition of return is largely automatic and by default disengages covert attention from each target shortly after the target has been acquired and attended to. An extension of the model (Itti *et al.*, 2003) adds a simple mechanism for the control of overt attention (eye movements) in rapidly changing environments. In this context, automatic inhibition of return is not desirable, as the slower oculomotor system cannot keep up with the ensuing rapid shifts of covert attention (up to 20 items/s). Thus, for tasks which require eye movement control rather than rapid covert search, we disable inhibition of return, so that covert attention locks onto the most salient location in the dynamic display. As this winning location may drift or even jump from one object to another according to the

dynamic video inputs, covert and resulting overt shifts of attention are still experienced, but typically at a slower pace which the oculomotor system can follow. In certain environments containing a distinctive actor or object moving against a fixed background, the model's behavior becomes dominated by tracking this actor or object. As suggested by our pilot human data in Section 8, it is probable that implementing some tradeoff mechanism between tracking and rapid attention shifts employing inhibition of return will be necessary to adequately model the perception of changing scenes in a videoclip. Bottom-up saliency is only one of the many factors which contribute to the guidance of attention and eye movements onto a visual scene. Attention indeed is complemented by rapid analysis of the scene's gist and layout to provide priors on where objects of current interest may be located, and facilitate their recognition (Oliva & Schyns, 1997; Henderson & Hollingworth, 1999; Hollingworth & Henderson, 1998; Rensink, 2000; Itti & Koch, 2001; Torralba, 2003). As in the VISIONS example, when specific objects are searched for, low-level visual processing can be biased not only by the gist (e.g., "outdoor suburban scene") but also for the features of that object (Moran & Desimone, 1985; Ito & Gilbert, 1999). This top-down modulation of bottom-up processing results in an ability to guide search towards targets of interest (Wolfe, 1994). Task affects eye movements (Yarbus, 1967), as do training and general expertise (Moreno et al., 2002; Savelsbergh et al., 2002; Nodine & Krupinski, 1998). Finally, eye movements from different observers exhibit different idiosyncrasies, which may result from possibly different internal world representations, different search strategies, and other factors (Hollingworth & Henderson, 2004a; 2004b).

This model has been restricted to the bottom-up guidance of attention towards candidate locations of interest. Thus, it does not address the issue of what might hold attention. A static fixation may last long enough for us to categorize an object or decide it not of interest. Return may then be inhibited until either fading memory or relations with another object yield further attention to it. Movement itself is a strong salience cue. For a moving object of interest, "what will it do next?" may be enough to hold attention, but once attention shifts, inhibition of return will then apply to the region of its imminent trajectory. Clearly, these general observations pose interesting challenges for both psychophysical and neurophysiological experiments and future modeling.

In most biological models of attention and visual search, such as the "Guided Search" model (Wolfe, 1994), the interaction between top-down commands derived from symbolic reasoning and low-level vision have been restricted to two simple effects: feature-based biasing (e.g., boost neurons tuned to a feature of interest like vertical motion, throughout the visual field) and spatial biasing (e.g., boost neurons responding to a given spatial region) of the kind supported by monkey physiology (Treue & Martinez-Trujillo, 1999). However, several computer models have

attacked the problem of more complex top-down influences. We have already seen the hypothesis-driven strategies employed by the VISIONS system, based on knowledge stored in schemas about their relationship (spatial or hierarchical) with other schemas. The model of Ryback et al. (1998) stores and recognizes scenes using scanpaths (i.e., sequences of vectorial eye movements) learned for each scene or object to be recognized. When presented with a new image, the model attempts to replay one of its known scanpaths, and matches stored local features to those found in the image at each fixation (cf. Noton & Stark 1971). However, Arbib and Didday (1975) argued that scanpaths are generated online from visual memory stored as retinotopic arrays, an approach consistent with the rapid generation of successive fixations performed by the model of Figure 5.

Schill et al. (2001) employ an optimal strategy for deciding where next to shift attention (based on maximizing information gain). But such computer vision models (a) do not address the problem of handling scenes which are both dynamic and novel and (b) lack biological correlates. On the other hand, Dominey, Arbib and Joseph (1995) extended a model of the cooperation of frontal eye fields and superior colliculus in the production of saccadic eye movements to show how top-down influences could be learned, via reinforcement learning, which can bias the effect of salience. In each case, cerebral cortex projected to the basal ganglia (BG) which in turn modulated activity in superior colliculus. In one case, the projection from inferotemporal cortex to striatum was adapted to associate each specific visual pattern with a leftward or rightward bias on the winner-take-all selection of dynamic targets. In a second study, the order of presentation of targets yielded, via patterns of connections from prefrontal cortex to BG, to a bias on the order in which the targets would be attended to if they were later presented simultaneously. Turning from sequences of saccades to sequences of hand actions, Arbib (this volume, Section 2.1) and Arbib and Bota (this volume, Section 4) discuss the transition from single actions to overlearned sequences, and the transition from overlearned sequences to sequences as the context-dependent expression of hierarchical structures, respectively.

5. GOAL-DIRECTED AND ACTION-ORIENTED GUIDANCE OF ATTENTION

To start building a scalable model, we must address the key question “What draws the viewer's attention to a specific object or action, and then expands that attention to determine the minimal subscene containing it?” The unfolding of a description and the answering of a question require that attention be driven not solely by low-level salience but also by the search for objects (and actions) that are deemed relevant to the unfolding task. We are motivated here by Yarbus’s classic study (1967), which revealed subjects’ dramatically different patterns of eye movements when inspecting a given static scene under different task guidelines. In this spirit, we extend the Itti-Koch model (Figure 5) to a model of goal-directed and action-oriented guidance of attention in real-world static and

dynamic scenes. The idea is to only memorize objects/events determined to be task-relevant, developing the current minimal subscene in symbolic STM. We first present a general conceptual architecture to achieve this, then briefly examine experimental and simulation results obtained with a partial implementation (Navalpakkam & Itti, 2005).

5.1. The SVSS (Salience, Vision and Symbolic Schemas) model

Our present conceptual model (Figure 6), the SVSS (Salience, Vision and Symbolic Schemas) model, uses a *Task Relevance Map (TRM)* to hold the locations currently deemed relevant to the task requirements. The TRM is a topographic map which acts as a top-down mask or filter applied to bottom-up activation. Thus, the attractiveness of a bottom-up salient region may be reduced if that region occurs at a location expected top-down to be of little behavioral relevance. Conversely, sensitivity to bottom-up salience is increased in visual regions expected to contain relevant targets. An initial implementation of some of the concepts outlined here will be described in Section 5.2, which focuses mostly on describing a new scene and less on answering complex questions about a scene.

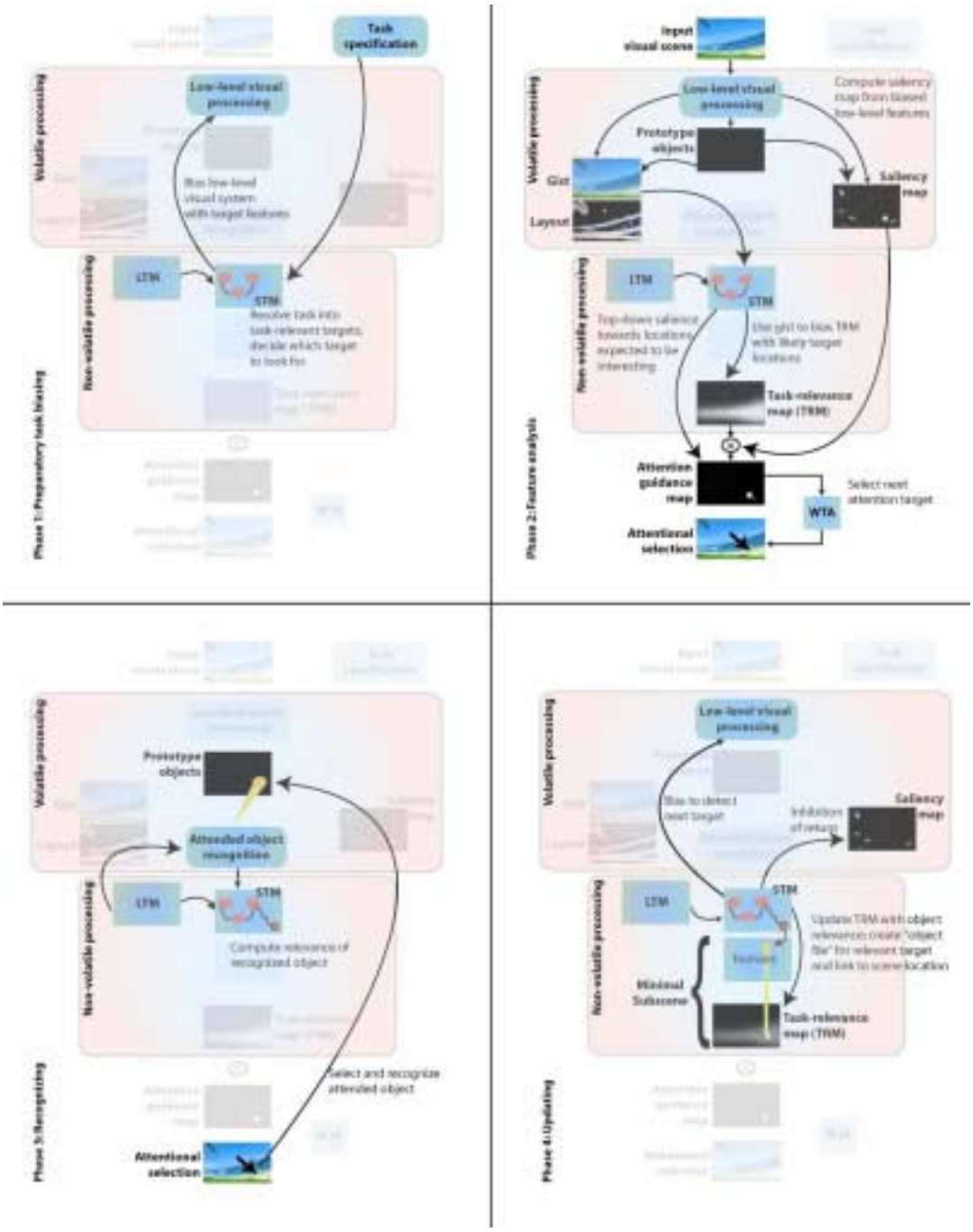


Figure 6: The SVSS (Salience, Vision and Symbolic Schemas) system for top-down (task-dependent) attention to objects in a static visual scene. Each cycle of activity in this model passes through 4 phases: *Preparatory task biasing* (preparatory top-down biasing of low-level vision once the task is known even before visual input is received); *Feature Analysis* (extraction of low-level visual features that will serve in computing gist, salience maps, and task-relevance maps); *Recognizing* (selecting the most salient and relevant location in the scene, attempting to recognize it, and updating the short-term memory (STM) based on how it relates to currently relevant entities); and *Updating* (by which recognized objects, actors or actions on a given attention shift are used to update both STM and task-relevance map (TRM), so that preparation for the next shift of attention can begin and be dependent on the current partial evaluation of the minimal subscene). Volatile processing refers to low-level representations which are updated in real-time from visual inputs, while non-volatile processing operates at a slower time scale, with one update occurring only with each attention shift. (Modified from Navalpakkam & Itti, 2005.)

In Section 13 we will consider more fully the extension of this model to handle dynamic scenes (changing over time) and how it may be linked to processes for the production and understanding of language. Here we note an interesting point of terminology. Within the vision community, we tend to use the term "scene" to refer to a static visual display, whereas when we go to a play, a scene is rather an episode which involves a set of overlapping actions by a limited set of actors. In this section, we will use "scene" primarily in the first sense, but with the understanding that we are setting the stage (to continue with the theatrical setting) to consider scenes that extend in space and time. We are here entering the territory of Zacks and Tversky's (2001) analysis of event structure in perception and cognition. A scene/event/episode in this spatiotemporal sense, too, will be considered as made up of a set of minimal or anchored subscenes/subevents/subepisodes each focused on a single action or relationship, where now we think of an action as extended in time rather than frozen in a snapshot.

As described below, rapid preliminary analysis of gist and layout may provide a spatial prior as to the location of possible targets of interest, in the form of an initial TRM (e.g., if looking for people, and presented with a scene whose gist is a beach, focus on the sand areas rather than on the sky). As time passes by, more detailed focal information accumulated through attention and gaze shifts serves to update the TRM, and to link locations marked as highly relevant in the TRM to additional STM representations, such as schema instances that may hold the identity and some visual attributes of objects currently deemed relevant. In VISIONS, for example, activation of a roof schema instance may direct attention to the region below to check for wall features. The TRM is one of several core components of the minimal subscene representation: it describes the spatial emphasis an observer may cast onto selected regions of the scene, but it is not concerned with identity of objects, actors, and actions which may be found in these regions. This identity and visual attribute information is stored in "object files" (the term for minimalist short-term descriptions of a specific visual entity employed by Treisman & Gelade, 1980), or assemblages of activated schema instances that hold information in STM on the appearance, identity, pose, and other aspects of scene elements which are part of the minimal subscene (cf. the state of the Arbib-Didday slide box). The TRM may also invoke information on where to look next for objects whose salience is defined in relation to the objects already found and their dynamic properties; for example, when attending to a ball in flight, highlight in the TRM the location of the expected future endpoint of the ball's trajectory. In this scenario, as in the static example of the roof and walls, perceptual schemas triggered by visual inputs (the ball or the roof) may prime additional schemas (trajectory extrapolation schema to estimate the ball's endpoint, or house schema prompting where to look for

walls), resulting both in a spatial priming of the related scene locations in the TRM and possibly in priming of related low-level visual features.

Where VISIONS offers no principled model of scheduling of schema instances, SVSS links this scheduling to an account of top-down mechanisms of attention which interact with salience, developing a theme of the Arbib-Didday model. Where VISIONS links schema instances to retinotopically defined regions of the intermediate representation in STM, SVSS extends STM by including a *symbolic* component which links certain "symbolic schemas" (which may or may not correspond to items in the lexicon) to those more "visual" schemas in the task representation and the currently active set of minimal and anchored subscenes. There are two notions at work here.

a) The first is to emphasize that our perception of a scene may depend both on specifically visual relations and on more conceptual knowledge about objects and actions. In this way, LTM encodes a number of symbolic concepts and their possible inter-relationships that seem to operate at the posited level of symbolic STM – for example, that a man has two hands. This complements the more specifically VISIONS-style use of LTM which enables recognition of a hand (activation of a perceptual schema for hand) to prime a schema for recognizing a human, whereas activation of a schema for a human could activate two instances of the hand schema with appropriate spatial constraints for possible incorporation into the unfolding schema assemblage.

b) The second is to consider STM in this extended sense as the bridge between the richness of the visual representation it contains and two different entities: (i) the verbal description of the scene, which is what we emphasize here; and (ii) short-term memory (STM) which is believed to result from hippocampal "tagging" of certain episodes for consolidation for subsequent recall. This link to STM provides some of the motivation for our theory, but cannot be treated at any length in this Chapter. In the terminology of Arbib (this volume, Section 1.3), we may relate STM to the cognitive structures (Cognitive Form; schema assemblages) from which some aspects are selected for possible expression, while the symbolic component of STM underwrites the semantic structures (hierarchical constituents expressing objects, actions and relationships) which constitute a Semantic Form. A selection of the ideas in the Semantic Form must be expressed in words whose markings and ordering provide a "phonological" structure, the Phonological Form.

We shall later turn to results from a partial implementation (Navalpakkam & Itti, 2005) of the SVSS architecture, but first we outline the mode of operation of the general scheme. Here we focus on analysis of a single static visual input. Computation is initiated by Phase 1 below, then cycles through Phases 2, 3 and 4 until Phase 1 is reactivated to update the task:

Phase 1. Preparatory Task Biasing: In situations where a task is defined in advance, e.g., "Describe who is doing what and to whom in the following scene" or "What is Harry hammering in the following scene?", the task definition is encoded as entities in symbolic STM, using prior knowledge stored in long-term memory (LTM). For instance, "Who is doing what and to whom" would prime the STM for humans and actions involving humans; "What is Harry hammering" would prime the STM for Harry, hammers, hammering actions, and objects which can be hammered on, like nails. One possible strategy for priming STM would be to provide an assemblage of primed schema instances for those entities deemed relevant to the task, either because they have been explicitly mentioned in the task definition (e.g., Harry, hammering), or because they are known from the LTM to be associated with the explicitly mentioned entities (e.g., a hammer, a nail). The issue then is to determine whether these primed instances can be linked to specific regions of the intermediate database (the preprocessed visual input) and given confidence levels which pass some threshold for accepting that the schema instance provides an acceptable interpretation of the region. To prioritize the search for these entities, one may compute a relevance score for each; for example, looking first for a hammer, then checking that Harry is using it, finally determining the object of the hammering may be a reasonable prioritization to answer "what is Harry hammering?". Such relevance-based scoring may be implemented through stronger output to the attentional mechanism of the schema instances which represent the more relevant objects, actors, or actions. As such, it is to be distinguished from the confidence level. The former reflects the priority for checking out a hypothesis; the latter reflects the extent to which the hypothesis has been confirmed. In general, before the scene is shown, little or no prior knowledge of spatial localization is available as to where the entities that are currently relevant and held in STM will appear once the scene is shown; thus the TRM at this initial stage is generally uniform. In specific situations, some initial spatial bias in the TRM may, however, already be possible; for example if the visual stimulus is presented on a relatively small computer monitor, the TRM may already assign low relevance to regions outside the monitor's display area.

As we have seen, the model can, in preparation for the analysis of the visual inputs already prime STM to bias its saliency-based visual attention system for the learned low-level visual features of the most relevant entity, as stored in visual long-term memory. (We use the usual shorthand here, omitting the phrase "the computer representation of the neural code for" when we speak of keywords and entities, etc.) For example, if a hammer is currently the most relevant entity, knowing that hammers typically feature a thin elongated handle may facilitate focusing attention towards a hammer, simply by enhancing the salience of oriented line segments of various orientations in bottom-up processing, while toning down non-diagnostic features, for example color which may vary from hammer to hammer.

In summary, we here propose a mechanism where prior knowledge stored in long-term memory may combine with task definition so as to populate the STM with a prioritized collection of task-relevant targets and possibly how they are related to each other. Next, the STM determines the current most task-relevant target as the desired target. To detect the desired target in the scene, the learned visual representation of the target is recalled from LTM and biases the low-level visual system with the target's features.

Phase 2. Feature Analysis: As the visual scene is presented and its gist is rapidly analyzed, the STM may impose additional biases onto the TRM, highlighting likely locations of the desired entity given the gist of the scene; for example, Harry and his hammer are more likely to be found near the ground plane than floating in the sky. The low-level visual system that is biased by the target's features computes the biased salience map. Cues from the TRM, together with the saliency map computed from biased low-level features (Phase 1), combine to yield the attention guidance map, which highlights spatial locations which are both salient and/or relevant. To select the focus of attention, we deploy a Winner-take-all competition that chooses the most active location in the attention-guidance map. It is important to note that there is no intelligence in this selection and all the intelligence of the model lies in STM and its deployment of schema instances on the basis of information stored in LTM. Given a current bottom-up input in the form of a saliency map biased to give higher weight to the features of desired objects, and a current top-down mask or filter in the form of a TRM, attention simply goes to the location where the combination of bottom-up and top-down inputs is maximized.

Regarding how bottom-up salience and top-down relevance may combine, it is interesting to note that non-salient locations may be relevant based on knowledge of spatial relationships between current objects of attention and desired relevant targets; for example, “where is John headed?”, “what is Mary looking at?”, “what is the hammer going to hit?”, or “where is that thrown ball going to land?”. We may term these special locations, which are of potential interest not because of any visual input but purely based on knowledge of spatial relationships, as “top-down salient”. Top-down salience complements bottom-up salience: a location is bottom-up salient if it has distinctive or conspicuous visual appearance, based on low-level visual analysis; conversely, a location is top-down salient if it is expected to be of interest, no matter what its visual appearance. Integrating top-down salience to the present framework may be achieved by allowing top-down salience to provide additive inputs to the TRM and resulting attention-guidance map, in addition to the bottom-up salience gated by relevance inputs described above. Top-down salience may result from the activation of expectation schemas; for example, a moving hammer may activate an expectation schema for what the object of the hammering may be, resulting in a spatial highlight in the

form of a cone of top-down salience below the hammer. Similarly, a flying ball may activate an expectation schema which will result in highlighting the expected endpoint of the ball.

Phase 3. Recognizing: Once a target is acquired, further processing is required to verify that the attended scene element meets the task requirements. This requires matching against stored LTM representations to recognize the entity at the focus of attention. Here, the low-level features or intermediary visual representations in the form of "prototype objects" (Rensink, 2000), which are intermediary between low-level features like edges or corners and full objects like a face, are bound together at the attended location to yield transiently coherent object representations (Treisman & Gelade, 1980; Rensink, 2000). The bound object representations are then analyzed for recognition, employing LTM to match the visual attributes of the bound representations to long-term memory traces of objects. If above-threshold confidence is reached that a given object or actor is present at the currently attended location, then the STM is called upon (in the next phase) to estimate the task-relevance and confidence level of the recognized entities, and to decide whether they are worth retaining or should be discarded; if no reliable recognition is achieved, one may either just ignore that location, mark it as being puzzling and as deserving further analysis later, or one may trigger other behaviors, such as slight head motion to change viewpoint, or an exploration of neighboring locations, in an attempt to achieve recognition.[2]

Phase 4. Updating: Recognition of an entity (object, actor, or action) as its confidence level passes some threshold serves to update the STM and TRM in two manners:

(i) If the recognized entity is found to be relevant (e.g., because it was one of the relevant entities already in working memory), its location is marked as relevant in the TRM, the entity is marked as found in the STM, and symbolic schemas may be activated to index key attributes of the entity in a form that may (but need not) link to language mechanisms. Which entity is next most relevant to look for is determined from the updated contents of STM.

(ii) If the recognized entity was irrelevant (e.g., either because it was unrelated to any of the entities in STM – as in building up a minimal or anchored subscene – or lacks independent interest to serve as a possible anchor for a new subscene), its location is inhibited in the TRM and its lowered activity in STM will reduce its likely impact on further processing.

---

[2] We trust that the reader has understood we are using the words "ignore" and "puzzling" and other such terms in this section as shorthand for detailed processes of computational analysis which in humans would usually be accomplished by the brain's neural networks without correlates in consciousness.

In this phase, the STM updates its state (e.g., if looking for a man's face, and a hand has been attended to, the STM may add a symbolic schema attesting that it has found a hand, which may help deciding where to look next to find the face by activating an expectation schema and associated top-down salience). In addition, the computed relevance of the currently attended entity may influence the system's behavior in several ways. For instance, it may affect the duration of fixation. In a last step, the STM inhibits the current focus of attention from continuously demanding attention (inhibition of return in the saliency map). Then, in preparation for a subsequent iteration of Phase 2, the visual features for the new most-relevant entity in STM are retrieved and used to bias the low-level visual system.

This completes one iteration, and each iteration involves one shift of attention. Subsequent shifts of attention will replay Phases 2-4 and incrementally build the TRM and populate the symbolic STM until the task is complete. Upon completion, the TRM shows all task- relevant locations and STM contains all task-relevant targets as high-confidence instances of perceptual and symbolic schemas, structured as one (or a concurrent set of) anchored subscene(s) representations, in a form appropriate to ground a variety of cognitive processes, including language production as in describing the scene or answering questions.

The STM and TRM are where SVSS, which is concerned with information processing from the retina to the minimal or anchored subscene, interfaces with models concerned with producing verbal descriptions of the scene from the subscene representations, for example using computational architectures which do for production what HEARSAY does for speech understanding. Although we have emphasized static scenes and object recognition in the above outline of SVSS, it is clear that it extends to recognition and linkage of agents, actions and objects that motivates this chapter. In this spatiotemporal extension, our subscene representation links a highly processed but still spatially structured visual representation of the visual world with a more symbolic aggregate of the recent past and expected future which provides one "now" that can be translated into language.

5.2. Implementing a Prototype

Our prototype software implementation of SVSS (Navalpakkam & Itti, 2005) emphasizes three aspects of biological vision: biasing attention for low-level visual features of desired targets, recognizing these targets using the same low-level features, and incrementally building a visual map of the task-relevance of every scene location. Task definitions are given to the system as unstructured lists of keywords; thus, there currently is no attempt in the prototype at parsing and exploiting the linguistic structure of more complex task definitions. Long-term memory is represented as an ontology in the sense used by the database research community, i.e., a collection of symbols

forming a graph where links represent several possible types of relationships between symbols. To each symbol is also attached visual description information, which allows both recognition of that entity in an image, and biasing of attention for the low-level features of the entity. Relationships currently implemented are: *is-a*, *includes*, *part-of*, *contains*, *similar-to*, and *related-to*. In addition, a link weight expresses the learned frequency of occurrence of a given relationship in the world; for example, a "pen" *is-a* "holdable-object" with a strength of 0.99 because nearly all pens are holdable, while a "car" *is-a* "holdable-object" with a strength of only 0.05 because most cars (except small toy cars) cannot be held. Thus, if looking for holdable objects, the system is more likely to be biased towards pens than towards cars. In the prototype, the ontology is small and hand-coded, with arbitrary weight values assigned to the various links in the graph. In future implementations, these weights would have to be learned, and the ontology possibly expanded, based on experience and training.

The STM of the prototype is initialized as a sub-ontology which is a copied portion of the long-term ontology, to include the concepts explicitly specified in the task definition as well as all related concepts down to a given weight threshold. Given this initial STM, a computation of relevance of the various entities to the task at hand gives rise to a single most-desired entity. Using visual attributes of entities stored in LTM, the visual attributes of the most-desired entity are retrieved. In the prototype, the visual properties of entities in LTM are encoded as a hierarchical collection of feature vectors, with one feature vector describing each learned view of an object (e.g., different photographs of a specific hammer), then combined to yield one feature vector for each object instance (a specific hammer), and finally for each object (a hammer). Feature vectors currently are very impoverished descriptions of each view, instance, or object, based on simple properties such as the amount (mean plus standard deviation) of various colors, various orientations, etc. typically observed for given objects. The prototype currently has no means of computing gist or top-down salience. Initially, hence, the TRM is initialized to a uniform unity value, and the low-level visual processing is biased for the low-level features of the most-desired target. This allows computation of the saliency map and guidance of attention to a given location. Recognition is based on matching the feature vector extracted at the attended location to its closest vector in the LTM. The recognized entity (if any passes the recognition confidence threshold) is then evaluated for relevance, using the LTM to evaluate how it may relate to the entities in STM and what the cumulative graph path weight is; for example, if the STM was interested in holdable objects and a car is found, the relevance of the car to the current STM can be computed by finding a path from car to holdable-object in the LTM graph, and computing how strong that path is. The TRM is then updated using the computed relevance value, which may either enhance a relevant location or suppress an irrelevant one. Finally, if the attended

object has a relevance score above threshold, its identity is remembered and attached to the corresponding location in the TRM. Thus, the output of the prototype is a TRM with a number of attached symbols at its most relevant locations.

This prototype presents obvious limitations, with its hand-coded ontology, weak object representation, and absence of gist and top-down salience. But it is a starting point in implementing the broader framework detailed above. The prototype was tested on three types of tasks: single-target detection in 343 natural and synthetic images, where biasing for the target accelerated target detection over two-fold on average compared to a naïve bottom-up attention model which simply selects objects in a scene in order of decreasing bottom-up salience; sequential multiple-target detection in 28 natural images, where biasing, recognition, working memory and long term memory contributed to rapidly finding all targets; and learning a map of likely locations of cars from a video clip filmed while driving on a highway. The model's performance on search for single features and feature conjunctions was shown to be consistent with existing psychophysical data. These results of our prototype suggest that it may provide a reasonable approximation to many brain processes involved in complex task-driven visual behaviors, and are described in details in (Navalpakkam & Itti, 2005).

6. EYE MOVEMENTS AND LANGUAGE: A BRIEF REVIEW

Before discussing explicit models which relate models like those of the previous two sections to language, we must note that, of course, we are not the first to investigate the relevance of visual attention to language. Henderson & Ferreira (2004) offer a collection of review articles that cover far more material than we can survey here. Instead we briefly note the contributions of Tanenhaus, Griffin and Bock. We note, without giving details, the related work of Tversky and Lee (1998) and Zacks and Tversky (2001) who explore "event structure" and the way events are structured spatially as a basis for relating that structure to language.

Griffin & Bock (2000) monitored the eye movements of speakers as they described black-and-white line drawings of simple transitive events with single sentences. However, the task was simpler than that we have set ourselves in that each drawing represented a single minimal subscene in which an agent acted in some way upon another agent or object. Eye movements indicated the temporal relationships among event apprehension (extracting a coarse understanding of the event as a whole), sentence formulation (the cognitive preparation of linguistic elements, including retrieving and arranging words), and speech execution (overt production). Their findings support the view that apprehension precedes formulation to provide the holistic representation that supports the creation of a sentence.

The experimental pictures depicted two types of events. Active events elicited predominantly active sentences in the experiment, regardless of which element was the agent. Passive-active events were those involving both a human and a non-human in which were typically described with active sentences if the human is the agent and with passive sentences if the human is the patient ("The mailman is being chased by the dog" and "The mailman is chasing the dog"). Their analysis included a variety of conditions. Here, we simply note that Griffin & Bock (2000) found an orderly linkage between successive fixations in viewing and word order in speech. Significant interactions between event roles and time period indicated that speakers spent significantly more time fixating agents before subject onset than during speech but spent more time on patients during speech than before subject onset. We would suggest that this means that the subjects create their STM of the scene, linked to the image, before they start to speak, and that in general – whatever the original sequence of fixations – the agent (or the human in the active-passive sentences) serves as anchor for the minimal subscene that will be described in the spoken sentence. Indeed, Griffin & Bock (2000) assert that analysis of passive-active pictures implies that speakers did not simply follow the causal structure of the events by fixating agents early and patients later. Rather, when patients were encoded as sentential subjects, they were fixated longer before subject onset than after whereas agents were fixated less before subject onset than during speech. Thus, the distribution of fixation times anticipated the order of mention regardless of sentence structure.

Griffin & Bock (2000) conclude that their evidence that apprehension preceded formulation supports the view that a holistic process of conceptualization sets the stage for the creation of a to-be-spoken sentence. Their data reflect only the most basic kind of sentence formulation, in English, involving minimal scenes and the production of single clauses. Nonetheless, the results point to a language production process that begins with apprehension of the structure of the scene and proceeds through incremental formulation of the sentence which expresses it.

Tanenhaus et al (2004) survey a rich set of experiments using eye movements to chart the role of referential domains in spoken language comprehension. The basic notion here is that a *referential domain* for a sentence provides the means to specify all the potential referents that satisfy the linguistic description. A definite noun phrase can then be used with multiple potential referents so long as the relevant domain defines a unique interpretation. For example, they observe that at a banquet you could ask the person sitting next to you to *please pass the red wine* even if there were six bottles of the same red wine on the table, but only one was clearly within reach of the addressee. Tanenhaus et al demonstrate that referential domains take into account behavioral goals expressed in a sentence. They do this by using the latency of eye movements probe the effects of having ambiguity or not in the possible

referents. They focus particularly on structural ambiguities which involve a choice between a syntactic structure in which the ambiguous phrase modifies a definite noun phrase and one in which it is a syntactic complement (argument) of a verb phrase. Just one example of these studies will have to suffice here. First note that *Pour the egg in the bowl over the flour* is temporarily ambiguous – it could be temporarily interpreted as *Pour the egg in the bowl* before it is discovered to be equivalent to the unambiguous instructions *Pour the egg that's in the bowl over the flour*. Chambers, Tanenhaus & Magnuson (2003) observed subject's eye movements in response to both these forms when subjects observed a variety of displays, each of which showed an egg in a cup, a broken egg (no shell) in a bowl, an empty bowl, and a pile of flour. In the "compatible competitor" case, the egg in the cup was also broken; in the "incompatible competitor" case, the egg in the cup was in its shell, unbroken. "Pour the egg" can only apply once the egg has been broken – thus "in the bowl" serves to disambiguate "the egg" in the compatible competitor case (liquid egg in glass) but not in the incompatible competitor case (solid egg in glass). In short, the subject entertains the idea that *in the bowl* in *Pour the egg in the bowl* as the Goal of the action if it seems unlikely that *in the bowl* could be an adjunct modifying the noun phrase, *the egg* because there is only one egg that is pourable.

When both potential referents matched the verb (e.g., the condition with two liquid eggs), there were few looks to the false goal (e.g., the bowl) and no differences between the ambiguous and unambiguous instructions. Thus, the prepositional phrase was correctly interpreted as a modifier. However, when the properties of only one of the potential referents matched the verb, participants were more likely to look to the competitor goal (the bowl) with the ambiguous instruction than with the unambiguous instruction. Thus, listeners misinterpreted the ambiguous prepositional phrase as introducing a Goal only when a single potential referent (the liquid egg) was compatible with a pouring action. In short, Chambers, Tanenhaus & Magnuson (2003) showed that the relevant referential domain defined can be dynamically updated based on the *action-based affordances* of objects.

By reviewing a number of other studies, Tanenhaus et al (2004) show that not only actions but also intentions, real-world knowledge, and mutual knowledge can circumscribe referential domains, and that these context-specific domains affect syntactic ambiguity resolution (among other processes). In many of the studies, the subject has had time to inspect the visual display before responding to a sentence or instruction. In these circumstances, the recording of the subject's eye movements supports the view that they map linguistic input onto action-based representations from the earliest moments of processing that input. Moreover, they note evidence that behavioral context, including attention and intention, affects even basic perceptual processes (e.g., Gandhi, Heeger, & Boyton, 1998; Colby & Goldberg, 1999) and that brain systems involved in perception and action are implicated in the

earliest moments of language processing (e.g., Pulvermüller, Härle, & Hummel, 2001). Given our concern in the Mirror System Hypothesis on the roots of language in manual actions and recognition of actions of the other, it is noteworthy that the final words of the Tanenhaus et al (2004) review are: "there is now substantial evidence that social pragmatic cues such as joint attention and intentionality are critical in early language development (e.g., Bloom, 1997; Sabbagh & Baldwin, 2001), as well as evidence showing that nonlinguistic gestures contribute to the understanding of speech (e.g., Goldin-Meadow, 1999; McNeill, 2000)."

7. LINKING SENSORIMOTOR SEQUENCES TO SENTENCES

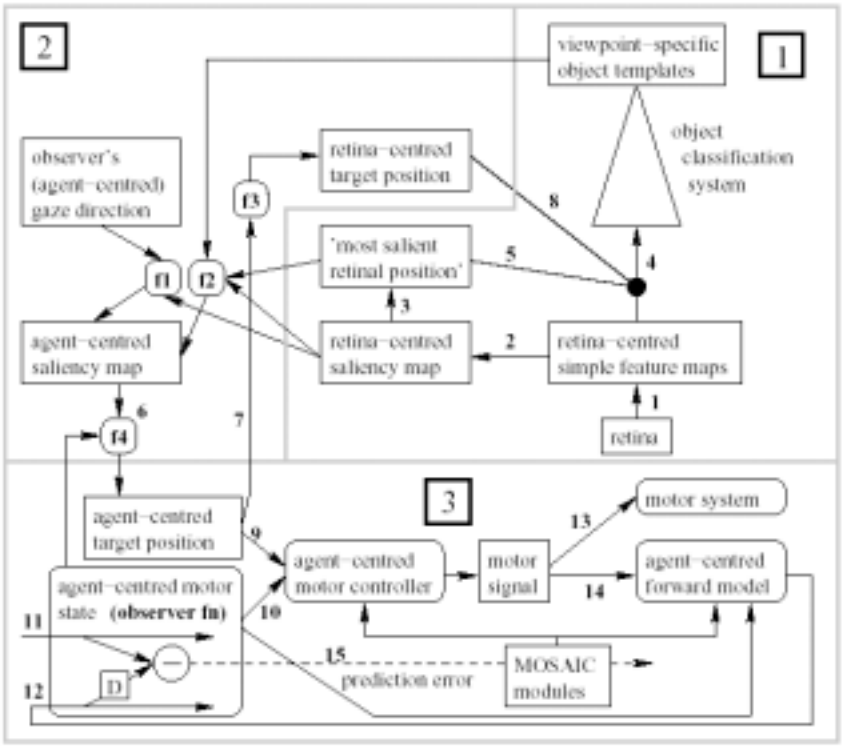


Figure 7: Sensorimotor circuit diagram (Knott, 2003) representing cognitive processes involved in executing a simple transitive action, such as reaching for a target object with the hand (functions in round boxes; data in square boxes.)

The theoretical framework which comes closest to what we seek in linking the SVSS (Salience, Vision and Symbolic Schemas) model of Section 5.1 is that offered by Knott (2003), which we briefly review. It builds on the Itti-Koch model (Section 4.2) and links this to a system for execution and observation of actions. Knott translates the sensorimotor sequence of attention to the scene into the operations involved in constructing the syntactic tree for its description. Box 1 of Figure 7 combines the Itti-Koch saliency model which selects an object for attention based on its saliency in the image with circuitry for classifying the selected object. Box 2 addresses how, once an object

becomes the center of attention, the agent's direction of gaze establishes a frame for further processing. However, Knott cites Tipper *et al.* (1992; 1998) concerning "action-centered"' representations in which objects compete in virtue of their closeness to the starting position of the hand which will reach for them. Thus, somewhat contrary to the actual labeling of the diagram, the key point is that determination of which is the most salient retinal stimulus may be gated by the position of the target object relative to the hand. Another possibility is that the observer can modulate the saliency map centered not on his own hand, but on the position of an *observed* agent or object (cf. Perrett *et al.,* 1989; Jellema *et al.,* 2000). We see that these generalizations are amenable to the processes offered by SVSS for building minimal and anchored subscenes.

Box 3 is a model of the motor controller which takes two inputs (a goal motor state and a current motor state) and generates a goal-directed motor signal. Knott (2003) emphasizes the MOSAIC model (Haruno, Wolpert and Kawato, 2001) for this subsystem, but this is not essential to the argument. For Knott, the goal input is the agent-centered position of the target and the current motor state is the agent-centered position of the agent's arm. Different controllers are required to execute different motor programs (e.g. *touch* vs. *push*) in relation to a target. This description assumes the agent is the observer but Knott notes the evidence for a mirror system – the mechanism by which we recognize actions in others using the same representations that are used to control our own actions – suggests the possibility of augmenting Box 3 by a network for "biological motion recognition".

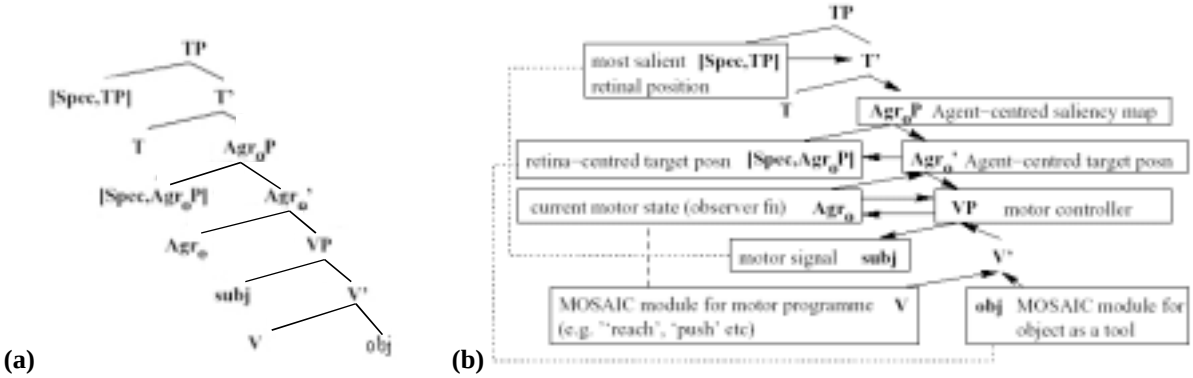


Figure 8: (a) Syntactic structure of a transitive clause. (b) The same structure with syntactic constituents associated with sensorimotor representations or sensorimotor functions. (Adapted from Knott, 2003)

We will say more below about the relation of Figure 7 to our own modeling, but here we note the key innovation of Knott's (2003) work, *relating the sensorimotor model to syntactic representations.* The idea is outlined in Figure 8. Knott assumes that the logical form of a sentence is a cognitive process rather than a representation of the world, and proposes that a syntactic constituent denotes an episode of activity within a sensorimotor system of the kind that

appears in Figure 7. The bottom right part of the syntactic tree of Figure 8a will be familiar to most readers: the idea that a structure (here called VP) links the subject subj to another structure (V') which combines the verb V and object obj. As we see in Figure 8b, Knott specifies linkages of these syntactic elements to processes in Figure 6.

Knott's analysis of the rest of Figure 7 incorporates ideas of Koopman and Sportiche (1991) and Pollock (1989) for generative syntax, but the syntactic complexities are outside the scope of this chapter. Suffice to say that these relate to the idea that the formation of a syntactic tree moves elements from one position to another but preserves links (traces) between the final position of an element and its initial position. This is exemplified by the notation "Who_i did John love [t_i]?" to indicate that the "Who" should be considered as having moved from the object position of "John loves X". The fact that the change from "loves" to "did … love" must also be explained gives a feel for why the complexity of the syntactic tree may be necessary even if at first sight it seem unduly complicated. In any case, Figure 8b embeds the basic VP tree in a larger tree which indicates (with dashed lines) the various movements that go into forming the final tree. Knott's aim is to characterize the notion of hierarchical position in a syntax tree in terms of the order in which representations become active during a sensorimotor action, with hierarchically high constituents becoming active before hierarchically lower ones.

The general idea, then, is the perception of a 'sentence-sized event' involves a sequence of transitions between different attentional states, each of which generates a distinctive side-effect in a medium for assembling linguistic representations. An agent is given a set of situations to observe, each accompanied by a sentence describing it, and must learn to generate appropriate sentences for similar situations. The agent processes each situation using a sensorimotor mechanism consisting of several different interacting components. Each of these components generates a side-effect of its operation at a linguistic level of representation. The architecture of the perceptual mechanism imposes a (partial) order on these side-effects, which can be construed as encoding certain aspects of the syntactic representation of the sentence to be expressed.

However, recall from the previous section the conclusion by Griffin & Bock (2000) that their evidence points to a language production process that begins with apprehension of the structure of the scene and proceeds through incremental formulation of the sentence which expresses it. Thus, using the terminology of the SVSS model (Figure 6) we would argue that Knott's theory is more pertinent if we build scene description on the state of the Symbolic Working Memory, with items tagged for relevance or "communicative salience", that is achieved following "apprehension of the subscene" rather than on the eye movements that went into the building of that representation. Competitive queuing (Bullock and Rhodes, 2003) would then yield the sequence of "virtual attention shifts" for this

internal representation. There will be further reordering as syntactic constraints re-order the words into accepted syntactic forms, much as Knott (2003) links sensorimotor processes to the "movements" engaged in formation of the syntactic tree in Figure 8.

Knott (2004) discusses how the model might be used to handle definite reference. During discourse, one maintains a record of all tokens that are understood to be accessible to both discoursers (but note the subtle view of referential domains offered by Tanenhaus et al., 2004, which goes well beyond a mere list of tokens). The perceptual process underlying the sentence *The man grabbed the cup* begins with an action of re-attention to an object already encountered and this is encoded as *the man*, not *a man*. This then triggers (in parallel) a representation of the man's local environment in a frame of reference centered on the man, and a mechanism for biological motion detection. These two events in turn jointly trigger identification of the action and identification of the target object. This corresponds to anchoring perception of a minimal subscene in an agent. We would also anchor it in a target object, but motion might be the salient cue that draws our attention to Harry's hand and thence to Harry and then to the theme of the motion – and the hand might not even be mentioned in the sentence "Harry is throwing a ball". But note that Knott is generating the sentence as attention shifts (like a scan-path) rather than after the scene-to-be-described has been chosen. Of course, one may also generate part of a sentence then update it as attention notes further aspects, or realizes they are needed for communicative clarity. In any case, our object is not to quibble with details of Knott's analysis, but rather to welcome it as a welcome signpost pointing us in the right direction for work integrating attention, scene description and action (Figure 7) with language (Figure 8).

8. DESCRIPTION OF DYNAMIC SCENES: A PILOT STUDY

The previous sections offer a conceptual framework within which to model how processes responsible for object recognition, action recognition, and rapid evaluation of the setting or gist of a scene integrate with top-down goals, task demands, and working memory in expanding a symbolic representation of a scene. However, almost all the efforts reviewed in the preceding two sections focus on the processing of a single, static scene whereas our eventual goal is to extend the study to dynamic scenes, in which change constantly occurs. To this end, we conducted a pilot study of eye movements in generating sentence after sentence in describing a rapidly changing scene. We had a subject (MAA) view 12 videoclips, and narrate a description as he observed each videoclip (each about 30s, divided into from 2 to 6 discrete scenes). Eye tracking apparatus allowed us to record his eye movements and superimpose them on a copy of the videoclip along with an audio recording of his narrative. Methods were as previously described (Itti, 2005). Briefly, the observer sat at a viewing distance of 80cm from a 22" computer monitor

(640x480 resolution, 60Hz double-scan refresh, 28x21deg field of view, mean screen luminance 30cd/m2, room 5cd/m2) and eye movements were recorded using a 240Hz infrared-video-based system (ISCAN RK-464) following a 9-point calibration procedure.

We analyzed these data to extract a range of hypotheses about the interaction between attention, object and action recognition, gist extraction, interplay between tracking and inhibition of return, and what we call "narrative momentum," in which not only does the speaker try to complete the description of the present subscene but may also direct attention to a subsequent minimal subscene whose description will extend the story in the description so far. In addition, below we show how such data may be more quantitatively analyzed so as to answer specific questions, for example, were objects and actors reported by the observer more bottom-up salient when the observer looked at them than objects and actors not reported?

The aim of our first broad and rather qualitative analysis of the data is to set targets for future modeling of a variety of constituent processes and to suggest strategies for the design of more structured tests of these models. Here we present some example sentences (in boldface) used to describe very different scenes and add observations made in relating the verbal description to the movement of the visual target on the dynamic scene:

Videoclip 1:

There's a document: Target traverses the field of view but this scene is too short to do any real reading. The subject notices someone moving across the scene, but despite this low-level salience, the sign holds his attention. **… it looks as if there is a riot … they've got a refrigerator or something:** People at left are the first target and anchor recognition of the gist of the scene as a riot. The refrigerator is then the most salient object – the eyes track the refrigerator in an attempt to make sense of what is being done with it. **Now a jeep or something comes in. A guy with a gun is moving across:** The eyes follows the jeep moving away but there is no time to describe this before a salient motion captures attention and anchors a new minimal subscene.

Successive subscenes may build on schemas and/or expectations of a previous subscene, or involve novel elements, when an unexpected object or action captures the attention. In addition, there is a real competition at the start of each scene between completing the ongoing description of the previous scene and starting the description of the new scene. There is a working memory issue here since the subject is processing a representation of the prior scene while building the representation of the new scene. Episodic memory is also at work since the subject can recognize a videoclip seen days or weeks previously.

Videoclip 2:

Now we're in front of the embassy with a couple of soldiers guarding it: The embassy sign is targeted then two policemen. The scene shifts to show soldiers. We hypothesize that the later sight of soldiers pre-empted the completion of the initial description, which thus conflated the policemen and the soldiers from two minimal subscenes.

Videoclip 3:

Looks like a snow scene with various rocks, a few trees. As we pan down it looks like it must be a hot pool because there's steam rising up ... This is one of the few cases where the motion of the camera is reflected in the verbal description. Interestingly, the steam is seen before the hot pool, but the description mentions the *inferred* pool before the steam that "implies" it. **And my goodness there's a monkey sitting in the pool so maybe it's in Sapporo or something – up in the island.** Bottom-up salience of a rather slight movement compared to the rest of the scene drew the subject's attention to the monkey. Note the long inference from "monkey in the snow" to the likely location. However, the subject mistakenly recalled the name of the town Sapporo where he should have named the island, Hokkaido and attempts repair with the last phrase.

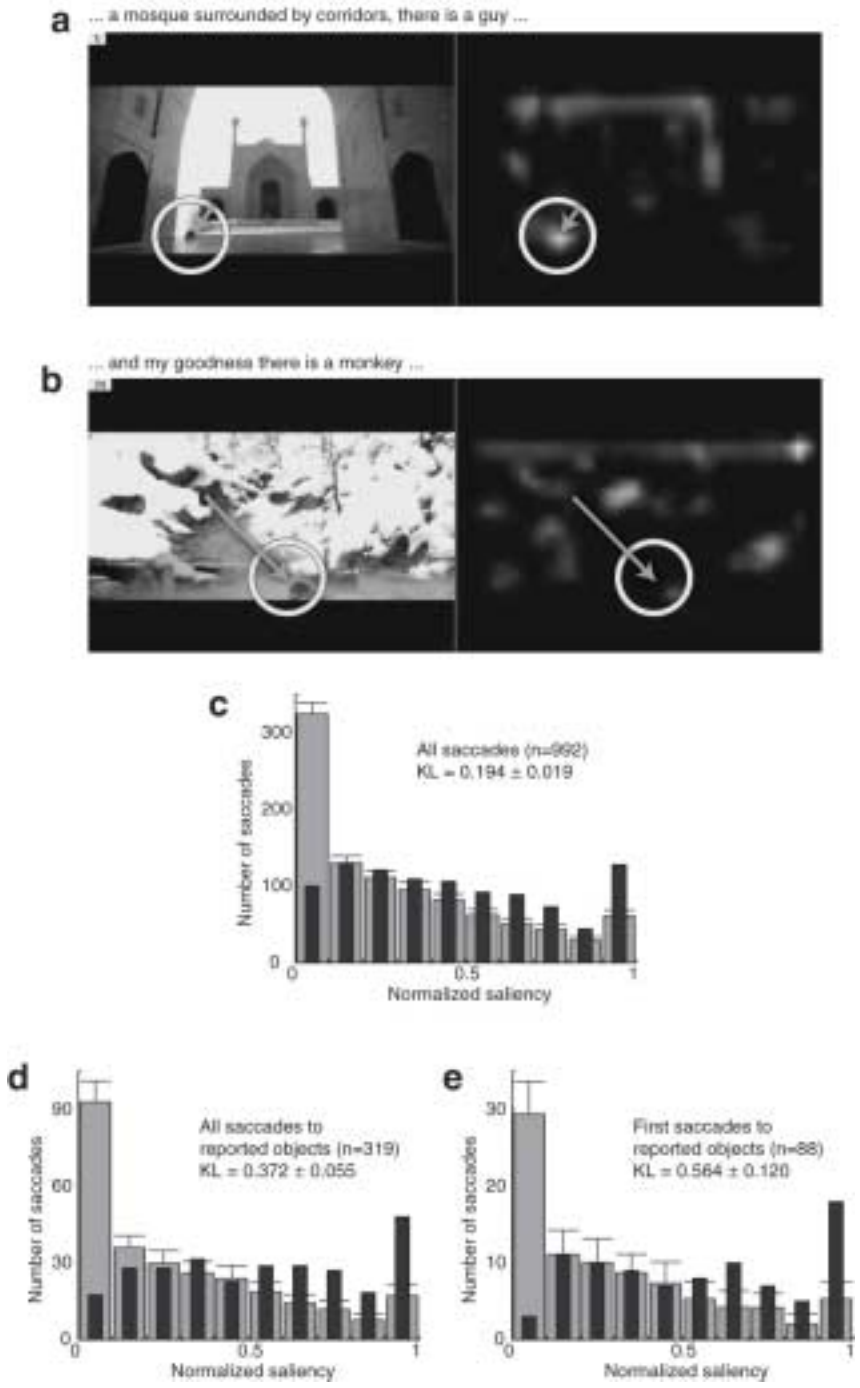


Figure 9. (a) Example human saccade, directed towards "a guy" reported verbally, who also was the most salient entity in the video clip at that instant. (b) Another example human saccade, directed towards "a monkey" also reported verbally, and who was not the most salient, although it was somewhat salient. (c)-(e) Histograms counting the number of human (thin dark bars) and random (wider light bars) saccades directed towards various salience values. Salience scores are derived from comparing the human and random histograms (see text for details), with a score of 0 indicating that humans did not orient towards model-predicted salient locations more than expected by chance, and scores above 0 indicating that bottom-up salience attracted humans. Scores are reported for all saccades (c), all saccades to reported objects (d), and the first saccades to reported objects (e), and indicate that out of all the locations gazed to by the observers, those which were reported verbally were significantly more salient than those which were not reported.

In a second analysis, we provide for the first time a quantitative answer to the simple question of whether objects and actors which are retained into the minimal subscene representation and which are reported by the observer tended to be more bottom-up salient than objects and actors not reported. To this end, we parsed the verbal reports in conjunction with the video clips onto which the scanpaths of the observer had been superimposed. This allowed us to mark those saccadic eye movements of the observer which were directed towards some objects, actors, or actions in the scene which were part of the verbal report. For example, as the observer described that "there is a monkey", we assigned a label of "monkey" to saccades directed towards the monkey in the corresponding video clip, and occurring within up to 5 seconds before or after the verbal report. To evaluate the bottom-up salience of saccade targets, we subsequently processed the video clips through the Itti-Koch model of bottom-up visual attention (Itti & Koch, 2000). At the onset of every saccade, we recorded the model-predicted salience at the future endpoint of the saccade, as well as at a random endpoint, for comparison (this random endpoint selection was repeated 100 times in our analysis, so that it provides a good online estimate of the overall sparseness of the dynamic saliency map at a given instant). For all saccades directed towards a reported scene element, and the first saccade directed towards each reported element, we compared the distribution of salience values at the endpoints of those saccades to what would have been observed by chance. We derived a score, which is zero if the targets of human saccades are not more salient than expected by chance, and greater than zero otherwise.

Figure 9(a) shows an example frame (left) and associated saliency map (right; brighter regions are more salient) of one clip, for the first saccade (arrow) directed towards "a man" also described in the narrative. The man was the most salient entity in the scene at that instant. Figure 9(b) shows another example from another movie clip, for the first saccade directed towards "a monkey", also reported in the narrative. In this example, the monkey is not the most salient item in the display, although it has non-zero salience.

Figure 9(c) shows the number of saccades histogrammed by normalized saliency at saccade targets, for human saccades (narrow dark bars) and random saccades (wider and lighter bars; S.D. from the 100-times repeated random sampling). Comparing the saliency at human saccade targets to that for random targets shows whether humans tended to orient towards regions with different saliency than expected by chance. Overall, saliency maps were quite sparse (many random saccades landing onto locations with near-zero saliency), but humans avoided non-salient regions (though fewer human saccades landing on these regions as compared to highly salient regions), while humans tended to select the few locations of near-unity saliency more often than expected by chance. The reported

KL score measures the difference between the human and random histograms. A KL score of 0.00 would indicate that humans were not attracted to salient image locations more than expected by chance. The KL score obtained here for all saccades is significantly higher than 0.00 (t-test, $p<10^{-27}$) as humans were more attracted to locations with above-average saliency than to those with below-average saliency.

Figure 9(d) shows a similar analysis to that presented in Figure 9(c), but only for those saccades that were directed towards some object, actor, or action reported in the narrative. The resulting KL score is higher than in Figure 9(c), indicating that items reported in the narrative had on average higher saliency than non-reported items (t-test, $p<10^{-13}$). Finally, Figure 9(e) shows a similar analysis, but considering only the first saccade towards reported objects, actors, and actions, that is, the initial noticing of these items. The KL score is even higher, indicating that when they were first noticed, items which would eventually be reported were more salient than other items (t-test, $p<0.0008$).

Thus, this pilot data (to be taken carefully as there is only one observer, although all results are statistically significant for that observer due to the reasonably large number of saccades) indicates that bottom-up salience may act as a facilitator, whereby more salient items actually not only attract gaze towards them, but are also more likely to be retained in the minimal subscene and verbally reported. Overall, these data demonstrated a rich variety of interactions between the factors driving attention (as reviewed above) and verbal reports. The merit of this first pilot study was thus to start cataloguing the interactions between visual scene understanding and language production when observers attempt to describe the minimal subscenes when they are exposed to complex scenes

We next conducted experiments with 32 clips filmed at a busy outdoors mall and a group of 9 observers, each asked post-hoc what they thought the minimal subscene was for each video clip. A working definition of the minimal subscene was provided and discussed with the subjects beforehand; subjects were instructed that the minimal subscene was the smallest set of objects, actors, and actions that are important in the scene, and those which they would want to mention when requested to give a short, one-sentence summary of what happened in each video clip.

Post-hoc rather than simultaneous reporting eliminated some of the factors of the previous experiments (e.g., narrative momentum, switching between subscenes, lacking verbal bandwidth to report some fixated events) – clearly both a plus and a minus, but indicative of how one may manipulate the video clips to focus on specific phenomena in each study. The study suggested the following broad categorization of factors influencing the selection of minimal subscenes:

(i) Factors derived from the gist, layout, and overall setting of the scene (e.g., objects belonging to the minimal subscene tended to be in the foreground, often occluding other objects; they usually occupied central positions within the scene's layout, and were present for extended time periods), including a number of extrinsic factors (the camera tended to follow the objects/actors/actions of the minimal subscene, often zoomed towards them);

(ii) bottom-up salience (e.g., isolated people, actors moving against a general flow, or brightly colored objects) with the noted caveat that, often, salient objects attracted attention although they were quickly discarded as being irrelevant to the current minimal subscene;

(iii) cultural and learned factors (e.g., finger pointing movements, facial expressions, postures), which made some events more relevant than others.

(iv) Personal preferences could guide different observers towards different minimal subscenes (e.g., a man playing with some electronic gadget attracted technology-oriented observers while a dog attracted animal-lover observers in the same clip). This second pilot study suggested stimulus manipulations by which some of these factors may be selectively emphasized or suppressed.

9. RECOGNITION OF EVENTS IN DYNAMICALLY CHANGING SCENES

In this section, we briefly note interesting efforts within AI on the recognition of events in dynamically changing scenes. A classic example of a scene analysis system, VITRA (Herzog et al. 1994), was able to generate real-time verbal commentaries while watching a televised soccer game. The low-level visual system recognizes and tracks all visible objects from video streams captured by a fixed overhead camera, and creates a geometric scene representation (the 22 players, the field and the goal locations). This representation is analyzed by series of Bayesian belief networks that incrementally recognize plans and intentions. The model includes a non-visual notion of salience which characterizes each recognized event on the basis of recency, frequency, complexity, importance for the game, and so on. The system finally generates a verbal commentary, which typically starts as soon as the beginning of an event has been recognized, but may be interrupted by new comments if highly salient events occur before the current sentence has been completed. However, VITRA is restricted to one highly structured environment and one specific task. Further, it is not a biologically-realistic model, and cannot scale to unconstrained environments as it constantly tracks all objects and attempts to recognize all known actions.

Another computer system for the cooperation between low-level perceptual analysis and symbolic representations and reasoning is provided by the work of Zhao and Nevatia (2004) on tracking multiple humans in

complex situations. A human's motion is decomposed into its global motion and limb motion. Multiple human objects in a scene are first segmented and their global motions tracked in 3D using ellipsoid human shape models. This approach is successful with a small number of people even when there occlusion, shadows or reflections are present. The activity of the humans (e.g., walking, running, standing) and 3D body postures are inferred using a prior locomotion model. Such analyses provide the basis for the hierarchical representation of events in video streams. To this end, Nevatia, Zhao and Hongeng (2003) developed an event ontology that represents complex spatio-temporal events common in the physical world by a composition of simpler events. The events are abstracted into three hierarchies. Primitive events are defined directly from the mobile object properties. Single-thread composite events are a number of primitive events with temporal sequencing. Multi-thread composite events are a number of single-thread events with temporal/spatial/logical relationships. This hierarchical event representation is the basis for their Event Recognition Language (ERL), which allows the users to define the events of interest conveniently without interacting with the low level processing in the program. For example, complex event "Contact1" is a linear sequence of three simple events: "approaching a person", "stopping at the person", and "turning around and leaving." Similarly, complex event "Passing_by" is a linear occurrence of "approaching a person," and "leaving" without stopping in between.

Such artificial intelligence systems pose the challenge of trying to better understand how the functional processes they embody may map onto brain areas and processes.

10. INVERSE MODELS, FORWARD MODELS, AND THE MIRROR SYSTEM

We here link the role of motor control in Figure 7 (Knott, 2003) to the more general notion of forward and inverse models, then point the reader to the discussion given by Oztop et al. (this volume) relating the FARS and MNS models of manual action and action recognition, respectively, to this general framework. We return to the distinction between the sign and the signified to distinguish between producing or perceiving a word and perceiving the concept that it signifies.

A *direct* or *forward model* of the effect of commands is a neural network that predicts the (neural code for) sensory effects in response to a wide set of (neurally encoded) commands. Conversely, an *inverse model* converts (the neural code for) a desired sensory situation into the code for a motor program that will generate the desired response. As described by Oztop et al. (this volume), Arbib & Rizzolatti (1997) used the notions of forward and inverse models to analyze the control of grasping:

1) The *execution system* leads from "view of object" via parietal area AIP (visual recognition of affordances – possible ways to grasp an object) and F5 (motor schemas) to the motor cortex which commands the grasp an observed object. This pathway (and the way in which prefrontal cortex may modulate it) implements an inverse model (mapping the desired sensory situation of having the object in one's grasp to the motor activity that will bring this about). This has been implemented in the FARS model (Fagg and Arbib, 1998; Arbib, this volume, Figure 4).

2) The *observation matching system* leads from "view of gesture" via gesture description (posited to be in superior temporal sulcus, STS) and gesture recognition (mirror neurons in F5 or area 7b) to a representation of the "command" for such a gesture (canonical neurons in F5) – another access to an inverse model. This has been implemented in the MNS model (Oztop and Arbib, 2002; Oztop et al., this volume, Figure 6). The *expectation system* is a *direct* model, transforming an F5 canonical command into the expected outcome of generating a given gesture. The latter path may provide visual feedback comparing "expected gesture" and "observed gesture" for monkey's self-generated movements, and also create expectations which enable the visual feedback loop to serve for learning an action through imitation of the actions of others.

Oztop et al. (this volume) stress that these inverse and forward models may in reality be seen as encompassing a whole family of inverse and forward models. Thus, in recognizing an action (as in the MNS model), we are not so much employing "the" inverse model, one of multiple inverse models best matching the observed interaction of hand and object. Rather, the system may recognize that the current action can better be viewed as a combination of actions already within the repertoire. Similar ideas have been applied in the "sensory predictor" hypothesis of mirror neurons which views *mental simulation* as the substrate for inferring others' intentions (Oztop et al., 2005).

The reader interested in this discussion of forward and inverse models should also consult the Chapter by Skipper et al. (this volume) which describes an active model of speech perception that involves mirror neurons as the basis for inverse and forward models used in the recognition of speech. In this view, both the facial and manual gestures that naturally accompany language play a role in language understanding.

11. MIRROR NEURONS AND THE STRUCTURE OF THE VERB

The MNS Model (briefly recalled in the previous section) provides the mechanisms whereby data on hand motion and object affordances are combined to identify the action. The claim is that, in concert with activity in region PF of the parietal lobe, the mirror neurons in F5 will encode the current action, whether executed or observed. However, F5 alone cannot provide the full neural representation of Grasp-A(Agent, Object), where Grasp-A is the current type of grasp. The full representation requires that the F5 mirror activity be bound to the IT activity

encoding the identity of the object and activity in STS (or elsewhere) encoding the identity of the agent, with "neural binding" linking these encodings to the appropriate roles in the action-object frame. This clearly involves a Working Memory (WM) that maintains and updates the relationships of agent, action and object. (Cf. our earlier critique of the approach of Baddeley, 2003; the discussion of relevant neurophysiology by Arbib & Bota in Chapter 5; and the review by Goldman-Rakic, Ó Scalaidhe and Chafee, 1999, of the variety of working memories implemented in macaque prefrontal cortex and with links to specific parietal circuitry.) This takes us from the hand-action focus of FARS-MNS to the study of actions more generally. We will build on the above to speculate about brain mechanisms involved in recognition of more general action-object frames ("The car is turning the corner") of the form Action(Agent, Object) and Action(Agent, Instrument, Object). We expect these to involve mechanisms far removed from the F5-Broca's mirror system for grasping, and must thus seek to understand how they nonetheless access Broca's area (and other language mechanisms) when such minimal subscenes are linked to language.

A note of caution: From a rigorous philosophical point of view, it is inappropriate to view a car as an agent and its turning as an action. Nonetheless, it can be useful to regard such verbs as "generalized actions". Indeed, linguists generally agree that most verbs should be given a syntactic analysis that includes thematic roles generalized from the (Agent, Instrument, Object) discussed above (Williams, 1995). Since our concern will be with nouns and verbs used in describing visual scenes, we have taken the pragmatic stance of deciding to use "action" for "that which a verb denotes" and "object" for "that which a noun denotes". Of course, this can only go so far, and the verb *to be* does not denote an action. Such subtleties are secondary to the present account of brain mechanisms linking attention, action and language.

With this, we devote the rest of the section to briefly noting the preliminary work by Vergnaud (personal communication) and Arbib seeking to gain deeper insights into the structure of the verb by giving a bipartite analysis of the verb in terms of mirror neurons and canonical neurons. Much work has been done by others on the lexicon, showing how a "lemma" –the "idea" of a word plus syntactic information (gender, number, …) – may be selected and then transformed into an articulatory code (see, e.g., Levelt, 2001, for an overview and Indefrey & Levelt, 2000, for a meta-analysis of relevant brain imaging). We will instead focus on how the linkage of a verb to its arguments may be illuminated by reference to the mirror system hypothesis (MSH). Briefly, we start from the observation that the two immediate arguments of the verb have different *linguistic* properties. The verb and its "object" can form a tight semantic and structural unit, unlike the verb and its "subject." There are also more abstract properties that distinguish "subjects" from "objects," relating to quantificational scope, anaphora, coreference,

control, weak cross-over, and incorporation (Kayne 1994). Such asymmetry is explicit in the verbs of languages such as Eastern Armenian (Megerdoomian, 2001) where the structure of an elementary clause is similar to that of, e.g., <Mary made honey drip.> where the verb now has two parts linked to the agent and object respectively, (Mary made) and (drip honey). The Vergnaud-Arbib hypothesis is that verbs should in fact receive this bipartite analysis even when it is not evident (as in most English sentences) on the surface. The resultant phrase structure

$[_S\ NP_S\ [_{vP}\ v\ [_{VP}\ V\ NP_O\ ]]]$

with v-V being the analysis of the verb in the sentence, is – we argue – the expression of a neural system whereby the human homologue of the F5 canonical system binding action to object underlies the merging of the V-component with the object NP_O; while the human homologue of the F5 mirror system binding agent to action underlies the merging of the subject NP_S with the complex v-VP.

We close by reiterating crucial distinctions emphasized in analysis (Arbib, this volume) of the Mirror System Hypothesis. We distinguish a mirror system for hand and face movements from mirror systems (posited but little studied to date) for other actions in the observer's own repertoire. Both of these are to be distinguished from mechanisms for the recognition of actions which are outside the observer's possible repertoire. MSH then hypothesizes that the production and recognition of words (phonological form) is based on an evolutionary refinement of the original mirror system for grasping, which has through evolution become multi-modal, integrating manual, facial and vocal gestures for communication. Thus it is only verbs related to manual action which are intimately linked to the networks that support the "semantic form" for the related action. For most nouns and other kinds of verbs, the recognition of the corresponding objects and actions must be conducted elsewhere and linked to the phonological form through processes which embody syntax to express perceived relations between actions and objects in a hierarchical structure which is expressed as a linear sequence of words.

12. CONSTRUCTION GRAMMAR

Both Knott and Vergnaud operate within the framework of generative grammar. Kemmerer (this volume) employs an alternative framework, construction grammar, in his work. We briefly recall some key ideas about construction grammar then suggest how "vision constructions" may synergize with "grammar constructions" in structuring the analysis of a scene in relation to the demands of scene description and question answering in a way which ties naturally into our concern with minimal and anchored subscenes.

In the next few paragraphs, we provide a brief exposition of construction grammar based on that provided by Croft and Cruse (2004, Chapter 9): Their starting point is to contrast the approach of generative grammar with the

use of *constructions* in the sense of traditional grammar. A basic example of such a construction is the *passive construction*:

a. The ball was kicked by Harry.

b. [Subject *be* Verb-PastParticiple *by* Oblique]

which combines the syntactic elements given in (b), including the subject noun phrase, the passive auxiliary verb *be* in some form, a past participle of a verb, and (optionally) a prepositional phrase with the preposition *by* and an oblique noun phrase. (A noun is said to be in oblique case when it is the predicate of a sentence or preposition but is not the subject of the sentence.) Semantically, the agent of the action in the passive construction is expressed by the object of the prepositional phrase, and the undergoer is expressed by the subject. Such a specific construction is to contrasted with a generative description which would seek to explain properties of a given construction in terms of general rules of the various components, with any idiosyncratic properties derived from the lexicon.

Generative grammar distinguishes the lexicon from the grammar, and this is seen as having three components – phonological, syntactic and semantic – with linking rules to map information from one component onto another. The "rule breaking" within any particular language is restricted to idiosyncrasies captured within the lexicon. Moreover, the rules inside each component are considered to be so highly intertwined and self-contained that each represents a separate structure that can be considered as being relatively autonomous. As Croft and Cruse (2004) note, Minimalist theory (Chomsky, 1995) recasts the phonological component as an "articulatory-perceptual interface" which links the language faculty to the perceptual-motor system and recasts the semantic component as a "conceptual-intentional interface" which links the language faculty to other human conceptual activity. The lexicon remains as the repository of idiosyncratic information, and as such provides information linking the three components together (Chomsky, 1995, pp.235-36). At first glance, the emphasis of Chomsky's Minimalism on an "articulatory-perceptual interface" and a "conceptual-intentional interface" seems compatible with our view of language within a broader framework action and perception. However, closer inspection shows that the Minimalist Program is far removed from a model of the speaker or hearer using language. The Minimalist Program characterizes which strings of lexical items are "grammatically correct" as follows (Figure 10): a set of lexical items is taken at random, the computational system then sees whether legal derivations can be built each of which combines all and only these elements. Spell-Out occurs when one of the legal derivations, if any, is chosen on the basis of some optimality criteria. The Computational System then transforms the result into two different forms, the Phonological Form, the actual sequence of sounds that constitutes the utterance, and the Logical Form, which

provides an abstract semantics of the sentence. There is no attempt here to model actual sentence production or perception – the process starts with words chosen at random and only at the end do we see whether or not they can be arranged in some way that yields a semantic structure. Of course, we have seen that Knott (Section 7) has offered a novel way of linking the formalism of generative grammar to sensorimotor structures. Thus while Figure 9 suggests that the Minimalist Program is at best indirectly related to perception and production, the possibility of using insights from generative grammar to develop an action-oriented approach to language has not been ruled out.

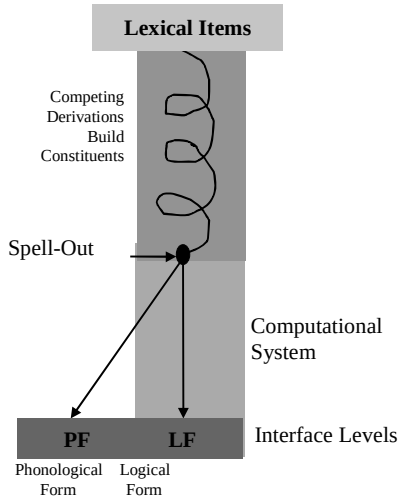


Figure 10. Derivations and the Computational System: The Minimalist Program (whose descriptive adequacy may be compared to Kepler's "conic sections" *description* of planetary motion). Contrast Figure 2 of Arbib (this volume) which places language within a broader framework action and perception (and whose adequacy may be compared to Newton's dynamical *explanation* of planetary motion).

However, construction grammar grew not out of a concern to model perception and production, but rather out of the need to find a place for idiomatic expressions like *kick the bucket*, *shoot the breeze*, *take the bull by the horns* or *climb the wall* in the speaker's knowledge of his language. Nunberg, Sag & Wasow (1994, pp.492-93) identified *conventionality* as a necessary feature of an idiom – its meaning or use requires an agreed upon convention and cannot be(entirely) predicted on the basis of its parts. But this suggests that the meaning of each idiom must thus be stored in each speaker's mind. Should we consider these meanings, then, as a supplement to the general rules of the syntactic and semantic components and their linking rules? In proposing the original construction grammar, Fillmore, Kay & O'Connor (1988) took a more radical step. Instead of adding idioms to the componential model, they suggested that the tools they used in analyzing idioms could form the basis for *construction grammar* as a new model of grammatical organization.

Croft and Cruse (2004) give a careful analysis of how a wide range of idioms can be captured by constructions to draw two general observations: (1) A given construction will often turn out to be just one of a family of related constructions. (2) The number and variety of constructions uncovered in studies of idioms imply that speakers possess a huge range of specialized knowledge that augments general rules of syntax and semantic interpretation on the one hand. They further note that many linguists working outside construction grammar have also examined schematic idioms and constructions and teased out the rule-governed and productive linguistic behaviors specific to each family of constructions. However, the key point is that constructions, like the lexical items in the lexicon, cut across the separate components of generative grammar to combine syntactic, semantic and even in some cases phonological information. The idea of construction grammar (Fillmore et al., 1988; Croft, 2001; Goldberg, 1995, 2003) is thus to abandon the search for separate rule systems within three separate components -- syntactic, semantic and phonological – and instead base the whole of grammar on the "cross-cutting" properties of constructions.

It is beyond the scope of this Chapter (and the present capability of the authors) to adjudicate on the relative merits of generative grammar or construction grammar, or to suggest whether and how one might integrate features of each, but we do want to offer certain comments relevant to the current enterprise. To this end, we note that Kemmerer (this volume) notes that even though *kick* is usually considered to be a prototypical transitive verb, it occurs in at least nine distinct active-voice constructions (Goldberg, 1995):

1. Bill kicked the ball.
2. Bill kicked the ball into the lake.
3. Bill kicked at the ball.
4. Bill kicked Bob the ball.
5. Bill kicked Bob black and blue.
6. Bill kicked Bob in the knee.
7. Bill kicked his foot against the chair.
8. Bill kicked his way through the crowd.
9. Horses kick.

These sentences describe very different kinds of events, and each argument structure construction here provides clausal patterns that are directly associated with specific patterns of meanings. We refer the reader to Kemmerer's chapter for his analysis of "action verbs, argument structure constructions, and the mirror neuron system." Here, we want to briefly argue that the approach to language via a large but finite inventory of constructions motivates the

return to visual scene interpretation armed with the notion that a large but finite inventory of "scene schemas" (from the visual end) may provide the linkage with constructions (from the language end) rich enough to encompass an exemplary set of questions we will ask and sentences subjects will generate. Each constituent which expands a "slot" within a scene schema or verbal construction may be seen as a hierarchical structure in which extended attention to a given component of the scene extends the complexity of the constituents in the parse tree of the sentence. This enforces the view that visual scene analysis must encompass a wide variety of basic "schema networks" in the system of high-level vision, akin to those relating *sky* and *roof*, or *roof*, *house* and *wall* in the VISIONS system (Section 3.2 above). Of course, we do not claim that all sentences are limited to descriptions of, or questions about, visual scenes, but we do suggest that understanding such descriptions and questions can ground an understanding of a wide range of language phenomena – see Arbib (this volume, Section 5.2) on "concrete sentences" versus "abstract sentences".

13. TOWARDS INTEGRATION

The detailed discussion of language processing is outside the scope of this chapter, but let us consider briefly the possible extension of the SVSS (Salience, Vision and Symbolic Schemas) model of Section 5.1 above into a complete model in which the more symbolic schemas of STM are linked to a system for language production, as schematized in Figure 11. As argued throughout this chapter, one of our main claims here is that the minimal or anchored subscene representation is the ideal interface between vision and language. In the case of meta-VISIONS, our top-level below the gist will be a spatial array of schemas. We suggest that the representation will not be a set of schemas so much as a graph of schemas. Thus there will be a schema for each recognized object, whereas an action will be a schema that links the schemas for agent and patient, etc., as constrained by the scene schemas ("vision constructions") of the previous section. This provides the connection to construction grammar, in that a scene schema may be instantiated with more or less links, acting in the top-down way in which the VISIONS house-schema sets up hypotheses on the presence of walls and windows – but here with the linkage being defined more by its extent in time than by its extent in space. Other links may correspond to spatial relations, giving the concrete grounding for prepositions, etc.

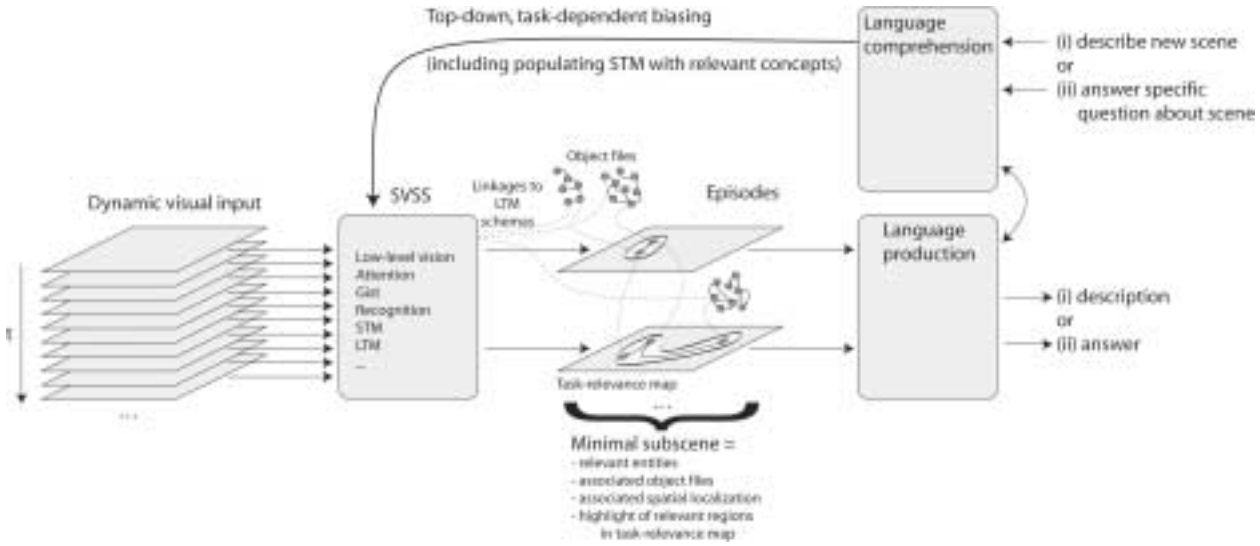


Figure 11. Minimal subscenes and symbolic schemas for episodes as the interface between vision and language.

The challenge for future research is to turn from the role of F5 and its homologues in recognition of manual actions to the study of the neural underpinnings of "general" action recognition (e.g., cars crashing). With this in place it should become possible to extend the Task Relevance Map (Section 5.1) to include actions and add them to the attentional specification. This takes us from keywords to relational structures. Thus Hit(∘,∘,∘) instructs the system to find any instance of hitting in the scene, whereas Hit(∘,ball,∘) instructs the system to find only those hitting-actions in which the instrument of the hitting is a ball. Similarly, the Task Relevance Map will not only link objects to locations, but will further provide links between objects representing any interactions between them. (Clearly, the same representation can accommodate spatial relations, such as a ball being on a table.)

Where the static version of the Itti-Koch model inhibits return of attention to an object for a short period after it is fixated, it is common in viewing a video clip (as in our pilot studies) to track an object until its interest declines or its role in a minimal scene is determined. Also, it is no longer enough to say where an object is in a scene; one must also give its trajectory – i.e., where it is at each time it is in the scene. However, the ensuing minimal scene will in general incorporate only general characteristics of the trajectory rather than its temporal details. Similarly, each action will have a temporal duration, where action recognition may now be based on the *time course* of the relationship between two objects.

In the case of HEARSAY, we have at the top-level a temporal sequence of sentences covering the single speech stream, and these in turn are linked to the sequence of words which best conform with that level. Sentences in turn link to semantic representations of whatever the sentence may mean, providing a basis for further behavior in

relation to the scene. Such a representation must include not just the relevant objects (nodes), and actions and relations (links) but a set of dynamic via points which segment the continuous visual input (as filtered through selective attention) into episodes which can be integrated into narrative memory. Our prototype (Section 5.2) uses classical Artificial Intelligence tools to implement the long-term and working memories of the model. These should be translated into schema-based representations, with possible neural correlates (Rolls & Arbib, 2003).

Such efforts must also factor into the extension of our Task Relevance Map model to build a graph (in the sense of nodes with some connecting edges) in which each minimal subscene becomes encapsulated as an "episode-node", and these nodes can be connected by links expressing salient spatial, temporal and causal relations. Such a model would enrich the very active study of episodic memory. We note studies which relate hippocampal activity to ongoing action (Lacquaniti et al., 1997; Elsner et al., 2002). The extension of the Task Relevance Map will address the diversity of hypothetical internal scene representations in the literature, such as the "world as an outside memory" hypothesis (O'Regan, 1992), the "coherence theory" according to which only one spatio-temporal structure can be represented at a time (Rensink, 2000), a limited representation of 5 to 6 "objects files" in visual short-term memory (Irwin & Andrews, 1996; Irwin & Zelinsky, 2002), and finally representations for many more previously attended objects in short-term and long-term memory (Hollingworth & Henderson, 2002).

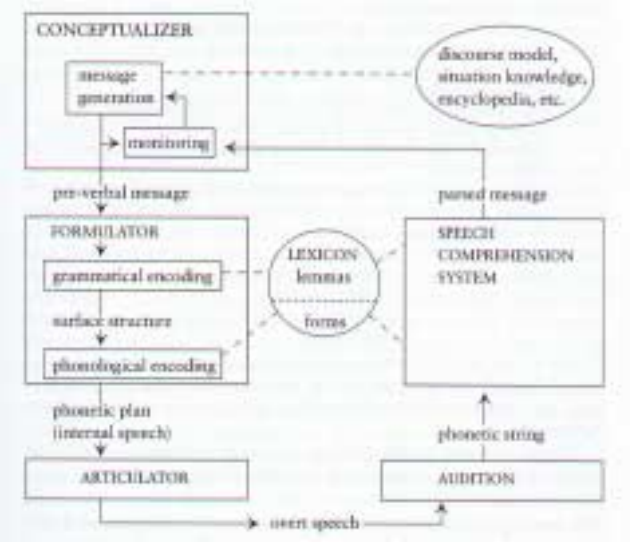


Figure 12: The left hand side of the figure shows Levelt's scheme for going from concept to phonological word in a scheme for the overall production of whole messages or sentences; at right is the complementary speech understanding system, for which HEARSAY provides our cooperative computation placeholder. (Adapted from W.J.M. Levelt, *Speaking*, MIT Press 1989, p.9.)

Figure 12 gives an overview of speech production and perception. Production of an utterance extends Levelt's scheme (Levelt, 1989; Levelt, Roelofs and Meyer, 1999; Levelt, 2001) for going from concept to phonological word; the complementary speech understanding system, for which HEARSAY provides our cooperative computation At the bottom level (not shown here), HEARSAY accesses the time-varying spectrogram of the speech signal. This is interesting because (a) the base level varies with time but, further (b) the representation at any time is a function of the recent past, not just the instantaneous present since a frequency estimate rests on a recent time window. Similarly, then, in developing our analysis of vision, we must note that the base level will include, e.g., MT activity, so we will have a spatial map of spatiotemporal jets (i.e., a collection of features centered on a particular point in space-time but extending in both space and time) as our base level. Note that this quickly gives way to quasi-symbolic representations. As we move up the hierarchy, we get symbolic representations (with confidence weights – a link to current treatments in terms of Bayesian models and Hidden Markov Models) that cover a greater and greater time span of lower-level representations raising the issue of the span of working memory at each of these levels.

In considering the implications for vision, note that object recognition can be thought of as based on a single time – but, of course, reflecting movement up the pyramid of spatial extent. However, when we look at action recognition, we carry this spatial processing into the temporal realm. In the MNS model, we think of the firing of a mirror neuron as in some sense a confidence level for the action it represents (leaving aside population coding for now), and that this level (akin to the lexical level in HEARSAY, for those words that are verbs) has an anticipatory component in the sense that the labeling of an action may change as it proceeds but will often be correct prior to completion of the action. Moreover, just as recognizing one word creates expectations about words that follow, so may perceiving one action create expectations of what further actions will unfold and towards what goal. Indeed, recent work in Parma reveals neurons that correlate with the next action of the sequence more than with the current action (Fogassi et al., 2005).

14. ACTION, PERCEPTION AND THE MINIMAL SUBSCENE

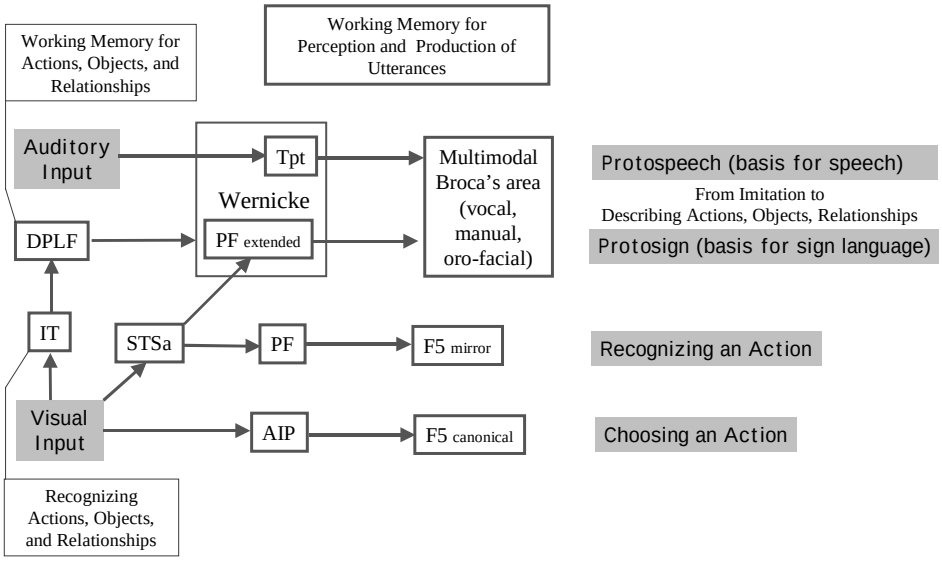


Figure 13. A high-level view of the cumulative emergence of three fronto-parietal systems: choosing a hand-related action, recognizing a hand-related action, and describing an action (in multiple modalities). This schematic builds on and modifies the schematic presented in Arbib (2001) to set goals for the modeling proposed here. See Arbib and Bota (this volume, Section 2.4) for further details.

Arbib and Bota (this volume, Section 6) built upon a schematic diagram (Figure 13) developed by Arbib & Bota (2003) to sketch the relations between action generation, action recognition and language within the framework of homologies between the brains of macaques and humans. They summarize the anatomical context for the generation and recognition of *single* actions and how these "lift" to communicative actions. Complementing this, this chapter has developed an overall framework for the functional analysis of how attention guides vision in the analysis of a dynamic visual scene, and how the structure of minimal subscenes or episodes (integrating the recognition of agents, actions and objects) may be linked with language production both in describing a scene or in answering questions about the scene.

Since the chapter has taken a long path through at times seemingly disparate elements, we review this framework by summarizing the key points of previous Sections. This summary complements the section by section overview of the Chapter given in Section 1.

1) A minimal subscene may be anchored by an agent, action or object and may then be extended to form an anchored subscene. A new focus of attention may (or may not) lead to the recognition of a new minimal/anchored subscene. In addition to overt shifts of attention to incorporate new objects and actions into a

minimal subscene, there will be "internal" shifts which attend to new features of an agent, action or object in fleshing out its description – as in going from "the man" to "the handsome old man on the left".

2) Gist and layout provide a framework for more detailed scene analysis, whether of a static or dynamic scene.
3) In linking vision and action to language, we will be particularly concerned with the relation between (i) building up a representation of minimal/anchored subscenes and scene description; and (ii) the role of top-down cues in visual attention involved in finding the appropriate subscene(s) on which to base the answer to a question.
4) We stress that the Mirror System Hypothesis (MSH) does not ask us to conflate the sign with the signified. Recognizing an action is part of recognizing a scene; uttering the verb to describe that action is itself a different action. MSH asks us to consider the parallels between the mechanisms for producing and recognizing manual actions and the mechanisms for producing and recognizing linguistic actions.
5) Within the general framework of schema theory (Figure 2; perceptual and motor schemas; schema assemblages; coordinated control programs) we considered both the VISIONS system (Figure 3a) which described how to represent a static visual scene by a hierarchically structured network of schema instances linked to regions of the image, and the HEARSAY system (Figure 3b) which described how to represent a speech stream by a hierarchically structured network of schema instances linked to time intervals of the spoken input. In each system, certain schema instances may be activated only to receive low confidence value; others may serve as "islands of reliability" which lead to activation of schemas instances which become part of the final interpretation of a region of the scene or an interval of the speech stream. Our challenge is to combine the insights of these two classical systems to described how to represent a dynamic visual scene (or, more generally, an integrated representation combining the analysis of multimodal sensory data with goals and plans for ongoing action) by a hierarchically structured network of schema instances each linked to a space-time region of the image. For example, a person may only be tracked over a certain time interval as they move from place to place; an action – whether observed or planned – will extend across a certain region of time but may be quite transient.
6) Since many motor schemas may initially be activated, constant processing is required to determine which actions are indeed executed – just as the "motor side" of eye control must execute a winner-take-all computation to select the next focus of attention.
7) In VISIONS, an intermediate database bridges between the image and the short term memory (STM, limited to the stock of schema instances linked to different regions of the scene). The state of this database depends both on visual input and top-down requests. As we extend our concern to dynamic scenes, we see that that the state of the

intermediate database will depend as much on the retention of earlier hypotheses as on the analysis of the current visual input. Indeed, much of that visual input will simply update the intermediate representation – e.g., adjusting the location of a segment or proto-object.

8) The SVSS (Salience, Vision and Symbolic Schemas) model explores the interaction of bottom-up salience and top-down hypotheses in building up a representation of a visual scene. Such interaction is crucial to understanding what "holds attention". Thus a static fixation may last long enough for us to categorize an object or decide it not of interest. Return may then be inhibited until either fading memory or relations with another object demands that it again receive attention. Movement itself is a strong salience cue. For a moving object of interest, "what will it do next?" may be enough to hold attention, but once attention shifts, inhibition of return will then apply to the region of its imminent trajectory.

9) In SVSS, visual processing is no longer neutral but will be influenced by the current task (e.g., answering a question about a specific object or action). The Task Relevance Map (TRM) holds the locations in the scene currently deemed relevant to the task requirements. We extend the VISIONS-like STM to include instances of more symbol-like schemas which provide a qualitative summary of just a few aspects for as long as the entity is relevant. In terms of producing a description of a scene, we envision STM as holding cognitive structures (Cognitive Form; schema assemblages) from which some aspects are selected for conversion into semantic structures (hierarchical constituents expressing objects, actions and relationships) which constitute a Semantic Form. Finally, ideas in the Semantic Form must be expressed in words whose markings and ordering provide a "phonological" structure, the Phonological Form.

10) Just as recognizing a visual scene involves much more than recognizing an action, so does recognizing a sentence involve much more than recognizing individual words.

11) A number of empirical studies have sought to link sentence perception and production to the eye movements of subjects looking at a visual display. However, these displays tend to be either formal arrangements of objects, or simple scenes which are exhausted by a single minimal subscene, thus avoiding a number of the important considerations which structure our analysis of complex scenes.

12) Knott has suggested that, in linking sensorimotor sequences to sentences, the sensorimotor sequence of attention to the scene (the scanpath) is translated directly into the operations involved in constructing the syntactic tree for its description. However, we suggest that Knott's theory is more pertinent if we build scene description on the state of the Symbolic Working Memory, with items tagged for the "communicative salience" that is achieved

following “apprehension of the subscene” rather than on the eye movements that went into the building of that representation.

13) MSH suggests a view of the verb as having two distinct parts which play different rules in the syntax of sentences: one part (cf. the canonical neurons of F5 and the FARS model) focuses on the linkage of an action to the object of that action; the other part (cf. the mirror neurons of F5 and the part of the MNS model complementary to the FARS model) provides the means for the linkage of an action to the agent of that action. Diverse circuitry for the recognition of agents, actions and objects must be linked to shared circuitry for the assemblage of words retrieved from the lexicon into well-formed sentences, and conversely for words to link to processes which modulate scene perception and ongoing action.

14) Construction grammar provides a framework for specifying a variety of constructions which provide syntactic rules for a number of different ways of linking agents, actions and objects. The approach to language via a large but finite inventory of constructions motivates a view of visual scene interpretation, compatible with the VISIONS framework, in which an inventory of “scene schemas” from the visual end may provide the linkage with constructions from the language end. Each constituent which expands a “slot” within a scene schema or verbal construction may be seen as a hierarchical structure in which extended attention to a given component of the scene extends the complexity of the constituents in the parse tree of the sentence.

15) Finally, we revisited SVSS and then used Levelt’s scheme for sentence production and understanding to provide the framework for the language processing system to be linked to systems for visual scene perception and action production.

MSH suggests that more than mere linkage of vision, action and language is at play here. It postulates an evolutionary progression from manual actions via complex imitation and pantomime to protosign and protospeech. On this account, certain neural components of the language system have their evolutionary roots in the praxic system – so that recognition and production of words is viewed as employing mechanisms homologous to those involved the recognition and production of manual actions. The attempt we have made here to unfold the interactions of task, attention and visual perception is part of the effort required to lift MSH from the production and recognition of single actions (Figure 13) to the understanding and planning of actions, agents and objects as defining minimal episodes and their integration into the overall episodes which structure our lives.

REFERENCES

Aboitiz, F., and García V., R., 1997, The evolutionary origin of the language areas in the human brain. A neuroanatomical perspective *Brain Research Reviews* 25: 381-396

Arbib, M. A., 1989, *The Metaphorical Brain 2: Neural Networks and Beyond*, New York: Wiley-Interscience

Arbib, M. A., and Didday, R. L., 1975, Eye-Movements and Visual Perception: A Two-Visual System Model, *Int. J. Man-Machine Studies*, 7:547-569.

Arbib, M. A., and Liaw, J. -S., 1995, Sensorimotor Transformations in the Worlds of Frogs and Robots, *Artificial Intelligence*, 72:53-79.

Arbib, M.A., 1981, Perceptual Structures and Distributed Motor Control, in *Handbook of Physiology, Section 2: The Nervous System, Vol. II, Motor Control, Part 1* (V. B. Brooks, Ed.), American Physiological Society, pp. 1449-1480.

Arbib, M.A., 2001, The Mirror System Hypothesis for the Language-Ready Brain, in *Computational Approaches to the Evolution of Language and Communication* (Angelo Cangelosi & Domenico Parisi, Eds.) Springer Verlag, Chapter 11, pp.229-254.

Arbib, M.A., and Bota, M., 2003, Language Evolution: Neural Homologies and Neuroinformatics, *Neural Networks* 16:1237–1260.

Arbib, M.A., and Caplan, D., 1979, Neurolinguistics must be Computational, *Behavioral and Brain Sciences* 2:449-483.

Arbib, M., and Rizzolatti, G., 1997, Neural expectations: a possible evolutionary path from manual skills to language, *Commun. Cognition*, 29:393-424.

Arbib, M.A., Érdi, P. and Szentágothai, J., 1998, *Neural Organization: Structure, Function, and Dynamics*, Cambridge, MA: The MIT Press (see Chapter 3).

Arkin, R.C., 1998, *Behavior-based Robotics*, MIT Press

Baddeley, A., 2003, Working Memory: Looking Back And Looking Forward, *Nature Reviews Neuroscience* 4:829-839.

Baddeley, A.D. and Hitch, G.J. (1974) Working memory. In *The Psychology of Learning and Motivation* (Bower, G.A., ed.), pp. 47–89, Academic Press

Ballard, D.H., Hayhoe, M., and Rao, P.P.R., 1997, Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20:723–767.

Bloom, P., 1997, Intentionality and word learning. *Trends in Cognitive Sciences*, 1:9-12.

Bullock, D., & Rhodes, B.J., 2003, Competitive queuing for planning and serial performance, in *The Handbook of Brain Theory and Neural Networks*, (M.A. Arbib, Ed.), Second Edition, Cambridge, MA: A Bradford Book/The MIT Press, pp. 241-248.

Cannon, MW, Fullenkamp, SC, 1991, Spatial interactions in apparent contrast: inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations, Vision Res, Vol. 31, No. 11, pp. 1985-98.

Chambers, C.G., Tanenhaus, M.K., & Magnuson, J.S., 2004, Actions and affordances in syntactic ambiguity resolution, J Exp Psychol Learn Mem Cogn. 30(3):687-96.

Chao, LL; Martin, A, 2000, Representation of manipulable man-made objects in the dorsal stream, *NeuroImage*, 12:478-484.

Colby, C.L. & Goldberg, M.E., 1999, Space and attention in parietal cortex. *Annual Review of Neuroscience*, 22:97-136.

Chomsky, Noam. 1995. *The minimalist program.* Cambridge, Mass.: MIT Press.

Croft, W. (2001) *Radical construction grammar.* Oxford: Oxford University Press.

Croft, W., and Cruse, D.A., 2004, *Cognitive Linguistics*, Cambridge: Cambridge university Press.

Dominey, P. F., and Arbib, M. A., 1992, A Cortico-Subcortical Model for Generation of Spatially Accurate Sequential Saccades, *Cerebral Cortex*, 2:153-175.

Dowty, D. (1991). Thematic proto-roles and argument selection. Language, 67(3), 547-619.

Draper, B.A., Collins, R.T., Brolio, J., Hanson, A.R., and Riseman, E.M., 1989, The Schema System, *International Journal of Computer Vision*, 2:209-250.

Elsner, B; Hommel, B; Mentschel, C; Drzezga, A; Prinz, W; Conrad, B; Siebner, H, 2002, Linking actions and their perceivable consequences in the human brain, *NeuroImage*, 17:364-372.

Emmorey, K. (2004) The Role of Broca's Area in Sign Language, to appear in "Broca's region" (Yosef Grodzinsky and Katrin Amunts, Editors), Oxford University Press, in press.

Epstein, R. and Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature, 392, 598-601.

Fagg, A. H., and Arbib, M. A., 1998, Modeling Parietal-Premotor Interactions in Primate Control of Grasping, *Neural Networks*, 11:1277-1303.

Fetz, F.E., and Shupe, L.E., 2003, Recurrent Networks: Neurophysiological Modeling, in: *The Handbook of Brain Theory and Neural Networks*, Second edition, (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, pp.960-963.

Fillmore, C.J., Kay, P., and O'Connor, M.K., 1988, Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language* 64:501-538.

Fogassi, L., Ferrari, P.F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G., 2005, Parietal Lobe: from Action Organization to Intention Understanding, *Science* (in press).

Gandhi, S.P., Heeger, M.J. & Boyton, G.M., 1998, Spatial attention affects brain activity in human primary visual cortex. *Proceedings of the National Academy of Science USA* 96:3314-3319.

Gibson, J.J. (1979) *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Goldberg, A., 1995, *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A., 2003, Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences, 7:*219-24.

Goldin-Meadow, S., 1999, The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3:419-429.

Goodale, M.A., A. D. Milner, L. S. Jakobson & D. P. Carey, 1991. A neurological dissociation between perceiving objects and grasping them. *Nature*, 349:154-156.

Gorniak, P., and Roy, D. (2004). Grounded Semantic Composition for Visual Scenes, *Journal of Artificial Intelligence Research*, 21:429-470. (a model of visually anchored grammar processing for referring expressions of visual scenes, tested extensively with natural data).

Grafton, ST; Fadiga, L; Arbib, MA; Rizzolatti, G, 1997, Premotor cortex activation during observation and naming of familiar tools, *NeuroImage*, 6:231-236.

Grezes, J; Armony, JL; Rowe, J; Passingham, RE, 2003, Activations related to "mirror" and "canonical" neurones in the human brain: an fMRI study, *NeuroImage*, 18(4):928-937.

Griffin, Z. and Bock, K. (2000). What the eyes say about speaking. Psychological science, 11, 274-279.

Groh, J. M., Born, R. T., & Newsome, W. T., 1997, How is a sensory map read out? Effects of microstimulation in visual area MT on saccades and smooth pursuit eye movements. *J. Neurosci., 17*(11), 4312-30.

Henderson, J. M., & Hollingworth, A. 1999. High-level scene perception. Annu Rev Psychol, 50, 243-271.

Henderson, J. M., & Hollingworth, A. 2003. Global trans-saccadic change blindness during scene perception. Psychol Sci, 14(5), 493-497.

Henderson, J.M., and Ferreira, F. (Eds.), 2004, *Interface of Language, Vision, and Action: Eye Movements and the Visual World*, New York, Hove: Psychology Press.

Herkovits, A. Spatial and temporal reasoning, Chapter 6, pp. 155-202, Kluwer Academic Publishers, 1997.

Herzog et al. 1994 (VITRA)

Hoff, B., and Arbib, M. A., (1993) Simulation of Interaction of Hand Transport and Preshape During Visually Guided Reaching to Perturbed Targets, *J . Motor Behav*. 25: 175-192.

Hollingworth, A., & Henderson, J. M. 1998. Does consistent scene context facilitate object perception? J Exp Psychol Gen, 127(4), 398-415.

Hollingworth, A., Henderson, J.M., 2002, Accurate visual memory for previously attended objects in natural scenes.

Indefrey, P. & Levelt, W. J. M., 2000, in *The New Cognitive Sciences*, ed. Gazzaniga, M. (MIT Press, Cambridge, MA).

Irwin, D.E. and Andrews, R., 1996, Integration and accumulation of information across saccadic eye movements. *Attention and performance XVI: Information integration in perception and communication*. Cambridge, MA: MIT Press, pp. 125-155.

Irwin, D.E. and Zelinsky, G.J., 2002, Eye movements and scene perception: Memory for things observed. *Perception and Psychophysics*, 64:882-895

Ito, M., & Gilbert, C. D. 1999. Attention modulates contextual influences in the primary visual cortex of alert monkeys. Neuron, 22(3), 593-604.

Itti, L., 2005, Quantifying the Contribution of Low-Level Saliency to Human Eye Movements in Dynamic Scenes, *Visual Cognition* (in press)

Itti, L., and Koch, C., 2000 A saliency-based search mechanism for overt and covert shifts of visual attention, Vision Research, 40:1489-1506.

Itti, L., and Koch, C., 2001, Computational Modeling of Visual Attention, *Nature Reviews Neuroscience*, 2:194-203.

Itti, L., Dhavale, N., and Pighin, F., 2003, Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention, In: *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*, pp. 64-78.

Itti, L., Gold, C., and Koch, C., 2001, Visual attention and target detection in cluttered natural scenes, *Optical Engineering*, 40(9):1784-1793

Jackendoff, R. (2002). *Foundations of language.* Oxford: Oxford University Press.

Jeannerod M., and Biguer, B., 1982, Visuomotor mechanisms in reaching within extra-personal space. In: *Advances in the Analysis of Visual Behavior* (D.J. Ingle, R.J.W. Mansfield and M.A. Goodale, Eds.), The MIT Press, pp.387-409.

Jeannerod, M., Arbib, M.A., Rizzolatti, G., and Sakata, H., 1995, Grasping objects: the cortical mechanisms of visuomotor transformation, *Trends in Neurosciences*, 18:314-320

Jellema, T., Baker, C., Wicker, B., and Perrett, D., 2000, Neural representation for the perception of the intentionality of actions. *Brain and Cognition*, 44:280–302.

Johnson-Laird, P, 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* Cambridge University Press.

Jordan, M.I., and Jacobs, R.A., 2003, Modular and Hierarchical Learning Systems, in: *The Handbook of Brain Theory and Neural Networks*, Second edition, (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, pp.669-672.

Kayne R.S., 1994, *The Antisymmetry of Syntax*. MIT Press, Cambridge, Mass.

Knott, A., 2003, Grounding syntactic representations in an architecture for sensorimotor control http://www.cs.otago.ac.nz/trseries/oucs-2003-04.pdf

Knott, A., 2004, Syntactic representations as side-effects of a sensorimotor mechanism Abstract for EvoLang 5, Leipzig, April, 2004

Koopman, H. and Sportiche, D., 1991, The position of subjects. *Lingua*, 85:211–258.

Krauzlis, R. J., Basso, M.A., & Wurtz, R.H., 1997, Shared motor error for multiple eye movements. *Science, 276*, 1693-1695.

Krauzlis, R.J. and Lisberger, S.G.,, 1994, A model of visually-guided smooth pursuit eye movements based on behavioral observations. *J. Comp. Neurosci. 1*, 265-283.

Kuczaj, S.A. (1982) On the nature of syntactic development, in *Language Development: Volume 1: Syntax and Semantics, (*S.A. Kuczaj, ed.) Lawrence Erlbaum Associates.

Lacquaniti, F; Perani, D; Guigon, E; Bettinardi, V; Carrozzo, M; Grassi, F; Rossetti, Y; Fazio, F, 1997, Visuomotor transformations for reaching to memorized targets: A PET study, *NeuroImage*, 5:129-146.

Lesser, V.R., Fennel, R.D., Erman, L.D., and Reddy, D.R., 1975, Organization of the HEARSAY-II speech understanding system, IEEE Trans. on Acoustics, Speech, and Signal Processing, 23:11-23.

Levelt, W.J.M., 1989, *Speaking*, Cambridge, M.A.: The MIT Press 1989.

Levelt, W.J.M, 2001, Spoken word production: A theory of lexical access, *Proc. Nat. Acad. Sci. (USA)*, 98:13464–13471.

Levelt, W.J.M., Roelofs, A., & Meyer, A.S., 1999, A theory of lexical access in speech production. *The Behavioral and Brain Sciences*, 22:1-75.

Lisberger, S. G., & Ferrera, V. P., 1997, Vector averaging for smooth pursuit eye movements initiated by two moving targets in monkeys. *J. Neurosci., 17*(19), 7490-502.

Luria, A.R., 1973, *The Working Brain*. Penguin Books.

Maratsos, M., and Chalkley, M. (1980) The internal language of children's syntax: the ontogenesis and representation of syntactic categories, *Children's Language*, vol.2 (K. Nelson, ed) Gardner Press.

Matthei, E. (1979) The acquisition of prenominal modifier sequences: stalking the second green ball, Ph.D. Dissertation, Dept. of Linguistics, University of Massachusetts at Amherst.

McCulloch, W.S., 1965, *Embodiments of Mind*, The MIT Press.

McNeill, D., Ed,, 2000, *Language and Gesture*. Cambridge, UK: Cambridge University Press.

Medendorp, WP, Goltz, HC, Vilis, T. & Crawford, JD, 2003, Gaze-Centered Updating of Visual Space in Human Parietal Cortex, *J. Neurosci*, 23:13

Megerdoomian K, 2001, Event structure and complex predicates in Persian, *Canadian Journal Of Linguistics-Revue Canadienne de Linguistique*, 46: 97.

Moran, J, & Desimone, R. 1985. Selective attention gates visual processing in the extrastriate cortex. *Science,* 229(4715), 782-4.

Moreno, F. J., Reina, R., Luis, V., & Sabido, R. 2002. Visual search strategies in experienced and inexperienced gymnastic coaches. *Percept Mot Skills*, 95(3 Pt 1), 901-902.

Navalpakkam, V., and Itti, I., 2005, Modeling the influence of task on attention, *Vision Research,* 45(2):205-231.

Nevatia, R., Zhao, T., and Hongeng, S., 2003, Hierarchical Language-based Representation of Events in Video Streams, IEEE Workshop on Event Mining, 2003.

Nodine, C F, & Krupinski, E A. 1998. Perceptual Skill, Radiology Expertise, and Visual Test Performance with NINA and WALDO. *Academic Radiology*, 5, 603-612.

Noton, D. & Stark, L., 1971, Scanpaths in Eye Movements during Pattern Perception, *Science (Washington),* 171, 308.

Nunberg, G., Sag, I.A., and Wasow, T., 1994., Idioms. *Language,* 70:491-538.

Oliva A. & Schyns P.G. (1997) Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology* 34(1) pp 72-107.

Oliva A. & Schyns P.G. (2000) Diagnostic colors mediate scene recognition. *Cognitive Psychology* 41(2) pp 176-210

O'Regan, J.K., 1992, Solving the "Real" Mysteries of Visual Perception: The World as an Outside Memory, *Can J Psych*, 46:461-488

Oztop, E. and Arbib, M.A., 2002, Schema Design and Implementation of the Grasp-Related Mirror Neuron System. *Biological Cybernetics*, 87: (2) 116-140.

Oztop, E., Wolpert, D., Kawato, M., 2005, Mental state inference using visual control parameters, *Cogn. Brain Res.*, 22:129-151.

Perrett, D., Harries, M., Bevan, R., Thomas, S., Benson, P., Mistlin, A., Chitty, A., Hiatenen, J., and Ortega, J., 1989, Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology*, 146:87–113.

Peters, A. (1985) The role of imitation in the syntactic development of a blind child. Paper presented at the Society for Research in Child Development, Toronto, April.

Pollock, J.-Y., 1989, Verb movement, universal grammar and the structure of IP. *Linguistic Inquiry*, 20(3):365–424.

Pulvermüller, F., Härle, M., & Hummel, F., 2001, Walking or talking? Behavioral and neurophysiological correlates of action verb processing. *Brain and Language*, 78:143-168.

Rensink RA. 2000 Seeing, sensing, and scrutinizing. *Vision Res*. 40:1469-87.

Rolls, E.T., and Arbib, M.A., 2003, Visual Scene Perception, in: *The Handbook of Brain Theory and Neural Networks*, Second edition, (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, pp.1210-1215.

Roy, D. (2002). Learning Words and Syntax for a Visual Description Task. *Computer Speech and Language*, 16(3). (visually-guided grammar acquisition for a visual scene description task; evaluated with human listening experiments)

Roy, D., and Mukherjee, N.. (In press). Visual Context Driven Semantic Priming of Speech Recognition and Understanding. Computer Speech and Language. (this paper is an audio-visually grounded implemented model of the Tanenhaus / Spivey online speech processing tasks in which visual attention interacts with speech).

Rybak, I A, Gusakova, V I, Golovan, A V, Podladchikova, L N, Shevtsova, N A, 1998, A model of attention-guided visual perception and recognition, *Vision Res*, 38: 2387-2400.

Sabbagh, M.A. & Baldwin, D.A., 2001, Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development*, 72:1054-1070.

Sablayrolles, P., 1995 Semantique formelle de l'expression du mouvement. De la semantique lexicale au calcul de la structure du discours en Francais. Ph.D. These, These IRIT, Université Paul Sabatier, Toulouse.

Savelsbergh, G. J., Williams, A. M., der Kamp, J. Van, & Ward, P. 2002. Visual search, anticipation and expertise in soccer goalkeepers. *J Sports Sci*, 20(3), 279-287.

Schill, K, Umkehrer, E, Beinlich, S, Krieger, G, Zetzsche, C, 2001, Analysis with saccadic eye movements: top-down and bottom-up modeling, Journal of Electronic Imaging.

Schirra, J., 1992, Connecting Visual and Verbal Space. In Proc of the 4th workshop on time, space, movement and spatio-temporal reasoning, Bonas, France.

Sillito, A M, Grieve, K L, Jones, H E, Cudeiro, J, Davis, J, 1995, Visual cortical mechanisms detecting focal orientation discontinuities, Nature, 378:492-6.

Simons, D. J. (2000). Attentional capture and inattentional blindness. *Trends in Cognitive Sciences*, 4, 147-155.

Tanenhaus, M. K., Chambers, C. G., & Hanna, J. E. (2004). Referential domains in spoken language comprehension: Using eye movements to bridge the product and action traditions. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press, pp.279-317.

Tipper, S., Howard, L., and Houghton, G., 1998, Action-based mechanisms of attention. *Philosophical Transactions of the Royal Society of London B*, 353:1385–1393.

Tipper, S., Lortie, C., and Baylis, G., 1992, Selective reaching: Evidence for action-centred attention. *Journal of Experimental Psychology: Human Perception and Performance*, 18:891–905.

Tomasello, M. (2003) *Constructing a language.* Cambridge, MA: Harvard University Press.

Torralba, A. 2003, Modeling global scene factors in attention. *J Opt Soc Am A Opt Image Sci Vis*, 20(7), 1407-1418.

Treue, S, Martinez Trujillo, J C, 1999, Feature-based attention influences motion processing gain in macaque visual cortex, *Nature*, 399:575-579.

Tversky, B., & Lee, P.U., 1998, How Space Structures Language. in *Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, (C. Freksa, C. Habel, K.F. Wender, Eds.): Lecture Notes in Computer Science, Volume 1404, Berlin, Heidelberg: Springer-Verlag GmbH, pp.157-175.

Ungerleider, L. G., and Mishkin, M., 1982, Two cortical visual systems, in *Analysis of Visual Behavior* (D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, Ed.), Cambridge, MA: The MIT Press.

Weymouth, T.E., (1986) Using object descriptions in a schema network for machine vision, Ph.D. Dissertation and COINS Technical Report 86-24, Department of Computer and Information Science, University of Massachusetts at Amherst.

Williams, E., 1995, Theta Theory, in *Government and Binding Theory and the Minimalist Program* (Webelhuth, G., Ed.). Oxford: Blackwell, pp.97-124.

Wilson, B., and Peters, A. (1984) "What are you cooking on a hot?": a blind child's "violation" of "universal" constraints. Paper presented at the Boston University Conference on child language, October.

Wolfe, J M, 1994, Guided search 2.0: a revised model of visual search, Psychonomic Bull Rev, 1:202-238.

Yarbus, A, 1967, *Eye Movements and Vision*, New York: Plenum Press.

Zacks, J. & Tversky, B. (2001). Event structure in perception and cognition. *Psychological Bulletin*, 127:3-21.

Zhao, T., and Nevatia, R., 2004, Tracking Multiple Humans in Complex Situations, *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 26:1208-1221.

Zwaan, R.A., and Radvansky G.A. , 1998, Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185.