A NEURAL NETWORK FOR SPATIAL RELATIONS: CONNECTING VISUAL SCENES TO LINGUISTIC DESCRIPTIONS

Lars Kopp

Lund University Cognitive Science
Kungshuset, Lundagård, S–222 22 LUND, Sweden
Lars.Kopp@fil.lu.se

This paper presents a system based on neural networks that can analyse spatial relations in a visual scene and connect them to appropriate linguistic descriptions. The system learns spatial concepts like “right of” and “above” by viewing a visual scene containing a number of objects and simultaneously receiving a text string describing the scene. The spatial relations between the objects in the scene are analysed with the aid of saccadic shifts of the focus attention. The system thus learns to correlate linguistic expressions for spatial relation with different kinds of saccades. After being trained, the system can correctly describe previously viewed scenes.

1. Introduction

One of the main questions in cognitive science is how different information codes interact in biological and artificial systems. In particular, the connection between the visual code and language is of central importance for understanding human cognition. The problem is that the information provided by a visual system must be translated into forms useable by a language capacity. In other words: How can we talk about what we see? (see Jackendoff, 1987).

In this paper, I will present a system based on neural networks that can analyse spatial relations in a visual scene and connect them to appropriate linguistic descriptions. The system learns spatial concepts like “right of” and “above” by viewing a visual scene containing a number of objects and simultaneously receiving a text string describing the scene. The spatial relations between the objects in the scene are analysed with the aid of saccadic shifts of the focus attention. The system thus learns to correlate linguistic expressions for spatial relation with different kinds of saccades. After being trained, the system can correctly describe previously viewed scenes.

2. Where and What in Vision

A visual scene, perceived by a human or an animal, is so complex that it is not possible to perceive the whole scene as one unit. Such a holistic perception would make the scene unique, and associations to other scenes and other perceptions would be impossible. It is therefore necessary to have a mechanism in the perceptual system that breaks down or fragments the scene to a more appropiate form of representation.

In order to solve the problem of representing complex visual scenes, I will in this paper focus on sequential representations. For example, if the scene contains two objects A and B standing in a certain spatial relation R to each other, this aspect of the scene could be represented by the sequence <A R B>.[1]

It has been found (Mishkin, Ungerleider and Macko, 1983; Zeki and Shipp, 1988) that humans and higher animals represent visual information in at least two important subsystems: the *where*- and the *what*-systems. The where-system only processes the *location* of the object in the scene. It does not represent the *kind* of object, e.g., whether it is an elephant or a mouse, but this is the task of the what-

1 The sequence should be read as a vector, not as a linguistic or symbolic representation.

Lund University Cognitive Studies – LUCS 32 1994. ISSN 1101–8453

system. The two systems work independently of each other and never converge to one common representation (Goldman-Rakic 1993). Physiologically, they are separated throughout the entire cortical process of visual analysis.

The where-system builds up what I will refer to here as a *spatial relation map*, where no information about the form of the object is represented. This means that the representation is a drastic generalization: an elephant is generalized to an undetermined object at a specific spatial position, indistinguishable from the representation of a mouse. This form of representation can be used for *variable binding* in collaboration with the what-system. The what-system represents *categories* of objects, without any information about their spatial location. An example of variable binding would be that an undetermined object in the where-system is bound to a representation of an elephant-object obtained from the what-system.

In natural environments, a significant problem is to *attend* to a stimulus of interest. The where-system is a part of the attention process, since the where-system supplies information about where to foveate in the scene. The fovea in the retina is exclusively concerned with form perception, not with the location of the objects in the scene.

The visual perception machinery also separates the information into several other domains such as *color, size, luminance, motion, depth* and *texture*. These kinds of representation are necessary for language understanding (Damasio and Damasio, 1990). For example, we can talk about color because color is represented as a separate domain and is very easy to attend to.[2] Lesions in area V4 of the visual cortex, where color is represented, result in color agnosia. Patients with this kind of brain damage cannot even imagine color. He or she is simply not conscious of the existence of color.

In the computer system that will be described here, representations are partitioned in a similar way. The system uses a very primitive representation of the category of objects, while it has a full representation of the spatial location of various objects. In the present implementation of the system, the color and other features of an object are not considered, but may be added in a future extension.

2 With few exceptions of languages which only have words for "light" and "dark", all languages contain basic color words (Berlin and Kay 1969).

3. Attention

When the brain processes a visual scene, some of the elements of the scene are put in focus by various attentional mechanisms (Posner, 1990). It is obvious that attention must be a very important property for identification and for learning in biological as well as artificial systems. Here, I will not address the general problems of attention in biological systems, but focus on attentional mechanisms necessary for grasping spatial relations.

In a natural scene, one of the basic problems is to locate and identify objects and their parts. However, in the artificial system studied here, the problem of attention is tremendously simplified since the representations of the objects and their locations are assumed to be known. The system can hence easily focus on one object at a time.

4. Saccades

When the brain analyses a visual scene, it must combine the representations obtained from different domains. One hypothesis underlying the simulations to be described in this paper is that attention shifts from domain to domain in a *sequential* way (see Crick, 1984). Since information about the form and other features of particular objects can be obtained only when the object is foveated, different objects can be attended to only through *saccadic* movements of the eye.

In the implemented model, the mechanism that generates attention to different objects and different domains results in a *sequence of readings of representations* from the different domains. They will be of the type "saccade, shape, color, saccade, shape …". For example, the processing of a visual scene consisting of a red car to the right of blue bicycle can be represented as "color: red, shape: car, saccade direction: left, color: blue, shape: bicycle." A sequence of readings produced by the system will here be called a *visual template*.

A common description of animal behavior is that an animal observes a stimulus, performs an action, and obtains a new stimulus, repeating the pattern over and over again. The coding of a visual scene in the simulated system uses a similar sequence "form-stimulus, saccade-action, new form-stimulus". However, the sequence is not executed concretely in the natural environment but instead with the aid of the internal representations.

Using the saccades as integrated parts of the processing sequences provides the the system with the possibility of representing abstract spatial concepts, for example "RIGHT_OF," with the aid of the

direction of the saccade. More precisely, it is the *internal action schema* for a saccade that is used to represent the spatial relation. The internal representation of the visual scene at a particular moment contains information about the focus of attention at that moment. In the system a saccadic movement to an object is randomly generated."Zero movement" is also an option which means that the attention remains fixed at the same object. For example, if the scene has three objects A, B and C, and the system attends to object C, it can perform three possible saccades: attend to object A, attend to object B, or continue to attend to C." The saccadic actions, together with readings of the features of the attended objects, will generate sequences of the kind described above.

5. Linguistic Correlation

The goal of the system that has been developed is to produce a linguistic description of a visual scene presented to the system. Like the visual templates, i.e., the sequences of readings from the scene, natural language is also sequential. The problem is to correlate the different visual templates, generated from the visual system, with the *language templates*, i.e., the text string presented to the language module of the system. In our simulations the linguistic units have been nouns and prepositional phrases.

The content of a visual template depends on the order of actions and their consequences. For example the sequence "shape: cat, saccade direction: down, shape: dog" will represent a scene where a cat is above a dog, while "shape: cat, saccade direction: up, shape: dog" represents the cat as being under the dog. The difference between the raw scene and the visual template representing it is that the sequential representation can synthesize different combinations of visual units into a template. Thus, the objects in a scene can be *grouped* in different ways. This "gestalt" property of the visual templates is necessary for higher cognitive tasks. In my opinion, sequential representations are therefore of great interest in biological or artificial systems.

The inputs to the system are visual scenes and linguistic templates, i.e., a text string. The visual scene can generate a large number of visual templates, depending on the order and direction of the saccades. The task of the system is to correlate different visual templates with the given text string. The correlation score of a visual template is determined from a matrix of numbers representing the probabalities of the co-occurence of a linguistic unit and a visual unit (saccade or object feature). The correlation value of a visual template is then sent as a multiplication or "resonance" factor to the correlation matrix. This multiplication represents a learning procedure, updating the co-occurence probabilities. The correlation matrix and the learning mechanism will be described in greater detail later.

During the learning phase of a simulation, the system thus learns to correlate linguistic units with appropriate units in the visual templates. After the learning phase, knowlege about the spatial concepts associated with the linguistic units can be used to let the system actively search information from the visual scene.

Suppose, for example, that two unidentified objects are shown next to each other in the scene. If the linguistic input to the system is the string "CAT RIGHT_OF DOG," the system can then find out which object is associated with the words CAT and DOG, since RIGHT_OF is known to correspond to a particular saccade direction.

After the learning phase the system can also perform the converse operation of generating a text which corresponds to a given scene. A visual scene, given as the sole input to the system, is not sufficient to determine a unique linguistic output, since the linguistic expressions describing spatial relation depend on the saccadic movements. Thus, different sequences of saccade sequences will generate different linguistic outputs. The different visual templates correspond to different ways of interpreting the scene.

6. Architecture of the System

I next turn to a description of the general architecture of the system used in the simulations. The system consists of several modules: The visual input is a *scene* consisting of a matrix of pixels. The *spatial relation map* is a matrix that identifies the positions of the objects (corresponding to the where-system of the brain). This information is extracted from the visual scene. The *pattern recognition network* classifies the visual patterns into objects (corresponding to the what-system of the brain).

In addition to the visual scene, a small section (5 by 5 pixels) is identified as the *focus of attention* (corresponding to foveating). Saccadic movements correspond to shifts of focus. The *saccade classifier network* sorts the relative movements into a small number of directions.

The *word classifier* breaks up the text input into units. The *correlation matrix* then connects units from the text input with the outputs from the pattern recognition network and the saccade

classifier.[3] Finally, in the case where the system is used to generate a description of a scene, this is handled by a *text output* module. The connections between the different modules are shown in figure 1. Each of them will be given a more detailed description below.

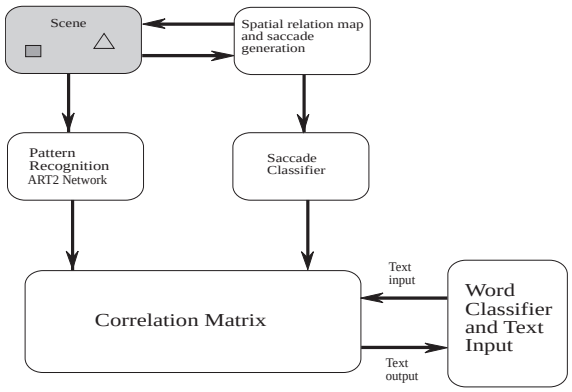

Figure 1: The general architecture of the system

7. General Description of the Implementation

Before I turn to the description of each of the modules, I want to outline the main computational structure of the system.[4] In brief, the system "looks" at the scene with its retina, which is directed to differents objects by a saccade generating unit. At each moment the system focuses on a small area of the scene. When the system focuses on an object, the information from that attended part of the scene will be directed to the pattern recognition network. This is an ART 2 network (Carpenter and Grossberg, 1987) that classfies the pattern as a particular type of object.

The saccade generating unit obtains possible saccade directions from the spatial relation map. In brief, it searches for areas of the scene where patterns that may represent objects are found. In other words, if an area of the scene is empty, it will never be focused upon. A random generator selects one of many possible directions to move from the present focus. The saccade motion is analysed in a prewired network which classifies the saccade as either stationary or as moving in one of four general directions.

The information about the objects obtained from the pattern recognition network together with the information from the saccade classifier is fed into a *relay station* which forms sequences of information units. It flips between the two domains so that saccades are intertwined with object information. A consequence of the random saccade process is that different possible visual templates, i.e., sequences of the type <object-shape(A,t), saccade(ΔP,t+1), object-shape(B,t+2), … >, are generated by the relay station.

The word classifier module is also an ART 2 network, providing a very elementary classification of the units from the text input. However, the network also functions as an output unit since activating a node may also activate a word. Thus the network can associate in both directions, and is able to generate text output, given input from the correlation matrix.

The correlation matrix correlates, unit by unit, the sequence of linguistic units, i.e., the language template, with the visual template, i.e., the sequence generated by the relay station. The number of units in the text string, determines the length of the correlating visual template. For example, a text input of the three linguistic units DOG RIGHT_OF CAT, is matched against the different visual templates containing three units that can be obtained from the sequence generated by the relay station. The system searches for the visual template that has the highest correlation with the linguistic template. The computational structure of the system is illustrated in some detail in figure 2.

8. Scene and Retina

Next, I turn to a more detailed description of the modules of the system. The scene in the system is a 64*64 matrix, where each pixel has 256 gray levels. An object can be placed at an arbitrary position in the scene by the user, by selecting a figure from a menu. The figure is represented in a predrawn 5*5 matrix, and can be dragged by the mouse from the menu to the scene.

When the system is focussing on an object in the scene, the image from the scene is copied to a "retina" also consisting of 5*5 pixels. In figure 3, two objects are present in the scene, and a 5*5 pixel retina is, at the moment, attending to one of the objects. The attention flips between the two objects, and the saccadic actions are transferred to the pattern recognition network one by one in a sequence. Before the image reaches the network, it is filtered to a courser scale, which reduces noise and increases the robustness of the pattern classification.

[3] As will be described below, there is no requirement that the visual template and the linguistic template contain the same number of elements.

[4] The program is written in C, and Microsofts QUICK-C compiler is used. The program runs on a PC with 1 Mbyte primary memory on a 486-Processor 66 MHz.

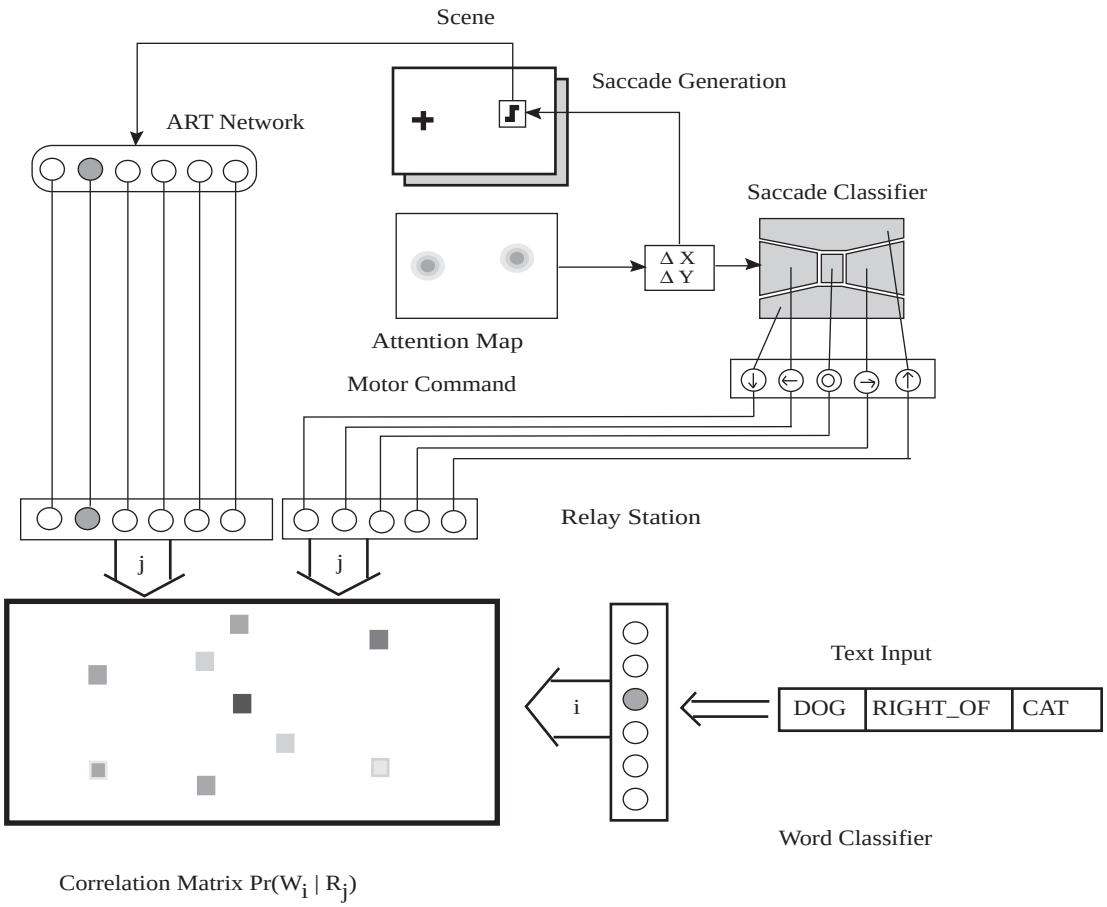

Figure 2: The computational components of the system.

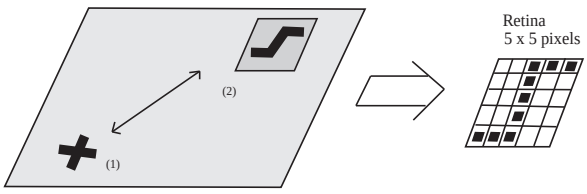

Figure 3: Scene with two objects and a 5*5 pixel retina.

9. Spatial Relation Map and Saccade Commands

A scene with two objects is shown to the left in figure 4. The objects in the scene are segmented by a Gaussian function into blobs, which are represented in the map to the right. A lateral inhibition mechanism finds the maxima, and results in a map of the maxima, the "spatial relation map," shown in figure 5.

The saccade system is then guided by the spatial relation map, which is used to generate appropriate saccades that shift the attention of the retina from one object to another. A random generator is used to select saccades in different directions. (See figure 6). The motor command unit, called saccade unit in figure 2, generates relative coordinates, which flips the attention by moving the retina from one object to another.

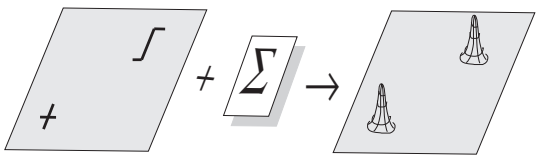

Figure 4: Generation of a spatial blob map.

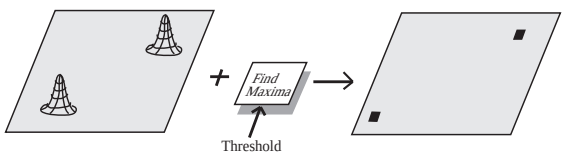

Figure 5: A sharp spatial relation map is created from the "blob map" by lateral inhibition.

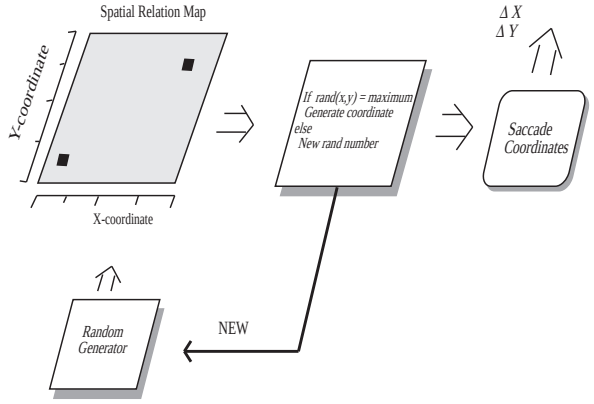

Figure 6: Motor command generation.

10. Saccade Classifier

The saccades generated by the saccade unit are then classified in the "saccade classifier". This classifier is prewired and cannot be modified by the system. The two dimensional space of X- and Y-coordinates is divided into five predefined categories, where each field is connected to a specific node, as shown in figure 7. In the figure the nodes are identified with the overall direction of the saccade. This division describes the receptive field of each of the five types of saccades. The central field corresponds to the zero saccade, i.e., no shift of attention.

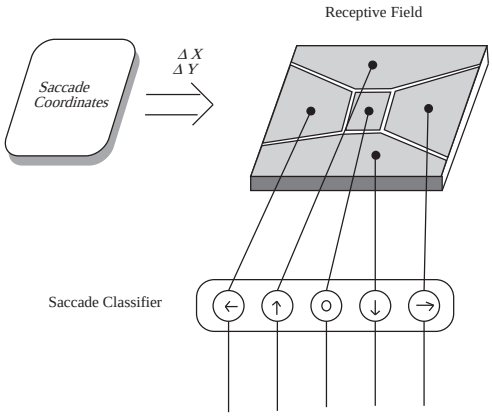

Figure 7: The saccade classifier.

11. Pattern Recognition ART Network

The pattern recognition unit is an ART 2 network (Carpenter and Grossberg, 1987) as illustrated in figure 8. The input to the network is lowpass filtered which has the effect of blurring the image. In lowpass filtering, high frequencies in the picture are suppressed and thus information is reduced. The vigilance factor in the ART network decides how many categories will be generated. The ART network is an unsupervised learning network that can automatically create new categories. Each picture on the retina is classified into an old or a new category by the network.

12. Relay Station

The relay station receives information from two channels: the pattern recognition network and the saccade classifier module, as shown in figure 9. The control unit of the relay station sequentially selects the information from the two channels and simply relays it to the correlation matrix. The control unit is preprogrammed to select the channels in a certain sequential order.

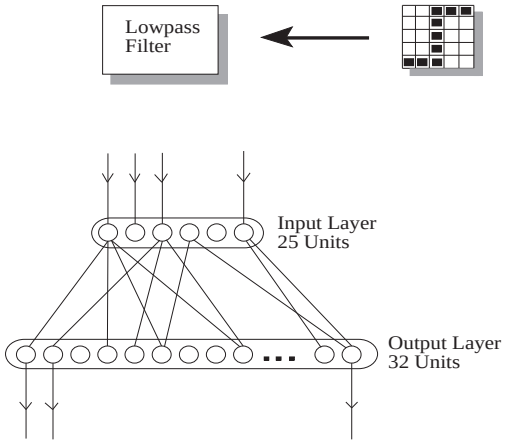

Figure 8: The pattern recognition ART2 network.

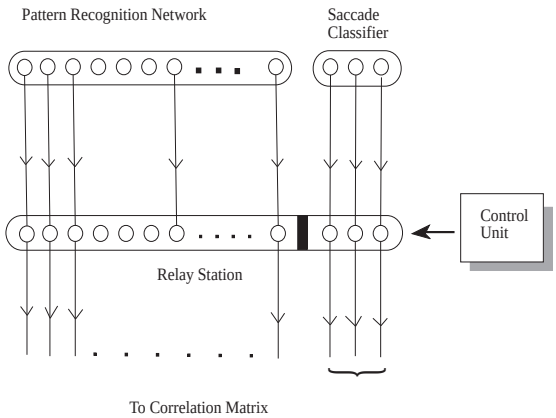

Figure 9: The relay station with two input sources: patterns and saccades.

13. Text Input and Word Classification

The system takes an ordinary text string as the input source to the word classification network. A word divider finds the words in a text string, which are stored in memory. This network simply creates a new node for each word. It also establishes a bidirectional connection between a word and the corresponding node so that a word in the input string activates a node, and an active node activates a corresponding word. The architecture of the word input unit is shown in figure 10.

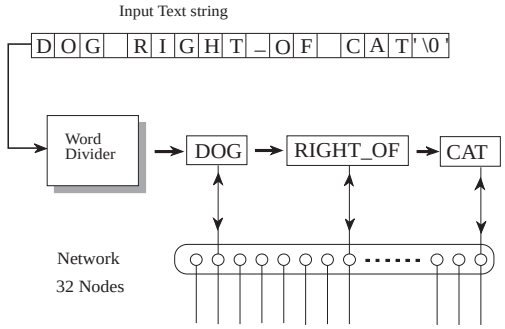

Figure 10: Text input and word classification network.

14. Correlation Matrix

The correlation matrix, illustrated in figure 11, receives a sequence of inputs from the word classifier network paired with a sequence of inputs from the relay station. These pairs of data are used to construct a probability matrix.

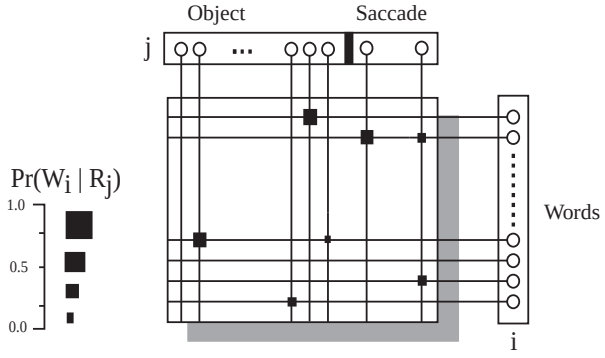

Figure 11: The correlation matrix

The pattern recognition network and the saccade classifier will, with the aid of the relay station, generate an alternating sequence of patterns and saccades of the form <R_1, R_2, R_3, R_4 ... > (the visual template). Similarly, the word recognition network generates a sequence <w_1, w_2, w_3, … w_n> of words from the input string(the linguistic template). These two sequences are then correlated in the following manner (figure 12): The word sequence <w_1, w_2, w_3, … w_n> is first matched with the sequence <R_1, R_2, R_3, R_4 ...> and the correlation corr(1) = $1/n*\Sigma_i Pr(w_i/R_i)$ is computed.[5] Here $Pr(w_i/R_i)$ represents the current value in the probability matrix of a word w_i occurring, given that the pattern or saccade R_i occurs.

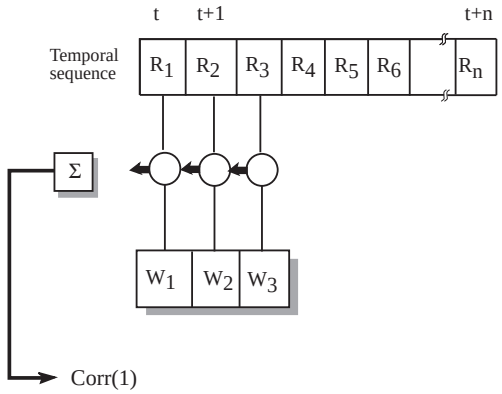

Figure 12: The first matching of the linguistic template to the visual template.

Next the word string is shifted one step to the right so that w_1 matches R_2 etc. and the correlation corr(2) = $1/n*\Sigma_i$ $Pr(w_i/R_{i+1})$ is computed. In this way, the word string is shifted stepwise to the right and the corresponding correlations corr(j) = $1/n*\Sigma_i$ $Pr(w_i/R_{i+j-1})$ are computed. This process is repeated for a fixed number of steps (standardly 100).

Suppose, for example, that the word string contains the three words DOG RIGHT_OF CAT and that the pattern saccade sequence begins with “shape: cat, saccade: right, shape: dog, saccade: left, shape: cat”. The first matching will be between the text string and “shape: cat, saccade: right, shape: dog.” Since Pr(DOG/shape: cat), Pr(RIGHT_OF/saccade: right), and Pr(CAT/shape: dog) should all be low, the total correlation will be low. In the next step the text string is matched against “saccade: right, shape: dog, saccade: left.” Also here the correlations are supposedly low, and thus the correlation is low again. However, in the third step, the text string is matched against “shape: dog, saccade: left, shape: cat.” Now the values of Pr(DOG/shape: dog), Pr(RIGHT_OF/ saccade: left), and Pr(CAT/shape: cat) are all very high and the result is a high correlation value.

The correlation values obtained in this way are continuously used to update the probability matrix.

[5] The index 1 in corr(1) indicates that the linguistic template is matched with the visual template from the first position of the visual template.

The total correlation value is used as a weight to change the single probability values in the matrix.[6] If the total correlation of the text string and the pattern sequence is high, the corresponding probabilities will increase, and if it is low, they will decrease. Initially, the probability values are set to small random numbers. During the training of the system, the conditional probabilities in the matrix will stabilize. If there are no ambiguous words, the probabilities will in general approach 1 or 0.

15. Training Phase

In order to facilitate learning, the system is often first trained on single words for objects. A scene consisting of just one object, say a dog-shape, is shown and the text input is just one word, say DOG. The sequence of patterns and saccade that is generated for this scene will consist of only two elements, namely shape: dog and saccade: zero. After updating the correlation matrix, the probability values Pr(DOG/shape: dog) and Pr(DOG/saccade: zero) will be comparatively large, while all other probabilities will be close to zero. This means that at this stage the system does not know whether DOG refers to a shape or a zero saccade.

If next a scene with just a cat-shape is shown and the lingustic input is just the word CAT, the system will learn that the word CAT is correlated with the cat-shape, while it will now assign both Pr(CAT/saccade: zero) and Pr(DOG/saccade: zero) very small values, and, as a consequence, the value Pr(DOG/shape: dog) will increase. In other words, the system will "realize" that CAT and DOG refer to shapes and not saccades.

After training the system with some shape-words, sentences containing spatial relations can then be used as the linguistic input. For example, suppose that the scene shows a fly shape vertically over a frog shape and the system has been trained to recognize the words FLY and FROG. If now the linguistic input is FLY ABOVE FROG, the system can learn to associate ABOVE with a downward saccade.

16. Test Phase

In the test phase the aim is to let the machine *describe* a simple scene. The input to the system is now only a visual scene containing known shapes, but where the spatial configuration of the shapes has not been shown during the learning phase. For an example see figure 13, which contains a "fly" shape, a "frog" shape and a "cat" shape.

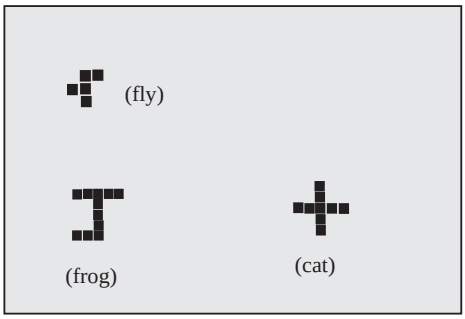

Figure 13: An example of a visual scene in the test phase.

With the aid of the saccade sequences generated from this scene, the system finds the linguistic template that has the highest correlation with the pattern-saccade sequence. Figure 14 illustrates how the association between a word and the output of the relay station is computed as a maximum of the conditional probabilities.

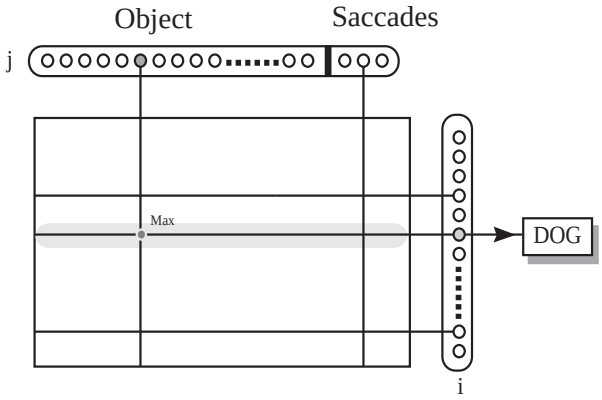

Figure 14: Generation of a new text string.

For the scene in figure 13, the system will generate strings like FROG UNDER FLY, FROG LEFT_OF CAT, but also longer sequences like CAT RIGHT_OF FROG UNDER FLY OVER FROG. Thus the system has no control for redundancies.

The system can also be shown to learn new shape words from the *context* of a scene. For example, suppose that the scene contains a dog shape above an unknown shape and another unknown shape to the right of the dog shape. If the corresponding linguistic input to the system is DOG ABOVE DUCK, the system will associate the word DUCK with the

[6] The actual calculations are rather complex and we will not present them here.

corresponding unknown shape and it will be able to use this knowledge about the object category in the analysis of a later scene.

17. Discussion

At the present stage of the system, it can only generate a stereotyped "language" as output during the test phase. The linguistic value of the system is thus limited. Another problem is that the system repeats the same text over and over again.

However several extensions of the system are possible to implement. One direction is to develop the system with scenes containing *moving* objects. For such scenes, one could train the system to associate appropriate motion types with words like TOWARDS, AWAY_FROM, HIT etc. Another extension is to introduce *properties* of the objects in the scene, for instance their colors. A system that contains more advanced visual information could also exploit the *depth* dimension of a scene, thus handling words like BEHIND.

As noted above, the linguistic output during the test phase is not filtered for repetitions. Another possible extension would therefore be to supply the system with various pragmatic filters on its linguistic output.

Finally, and perhaps most importantly, in the current version the shape recognition module is trivial. A more useful system would contain a realistic shape analysis based on video inputs from actual scenes. Such an extension would have immense potentials for technological applications.

Acknowledgements

An earlier version of the system was developed while the author was employed at the Cognitive Science Center, Roskilde University Center. The author wishes to thank professor Niels Ole Bernsen for proposing the project and for helpful ideas and discussions. The current version of the system was completed at Lund University Cognitive Science. Christian Balkenius, Peter Gärdenfors, and Robert Pallbo have all helped in various ways.

References

Carpenter, G. & Grossberg, S. (1987), "ART2: Self-organization of stable category recognition codes for analog input patterns", *Applied Optics*, 26, 4919–4930.

Crick, F. H. C. (1984),"Function of the thalamic reticular complex: The searchlight hypothesis", *Proceedings of the National Academy of Science USA,* 81, 4586–4590.

Crick, F. H. C. and Koch, C. (1990),"Towards a neurobiological theory of consciousness", *Seminars in the Neuroscience*, 2, 263–275.

Crick, F. H. C. (1979),"Thinking aboutthe brain", *Scientific American,* 241, 219–232.

Damasio, H. and A. Damasio (1990),"The neural basis of memory, language and behavioral guidance: advances with the lesion method in humans", *Seminars in the Neuroscience*, 2, 277–286.

Goldman-Rakic, R. S. (1993),"Dissociation of object and spatial processing domains in primate prefrontal cortex", *Science*, 260, 1955–1957.

Jackendoff, R. (1987),"On Beyond Zebra: The relation of linguistic and visual information," *Cognition,* 26, 89–114.

Posner, M. I., and Petersen S. E. (1990),"The attention system of the human brain", *Annual Review of Neuroscience,* 13, 25–42.

Mishkin M., Ungerleider, L. G. and Macko, K. A. (1983),"Object vision and spatial vision: Two cortical pathways", *Trends in Neuroscience*, 6, 414–417.

Zeki, S. and Shipp, S. (1988), *Nature,* 335, 311.