Dynamics of picture viewing and picture description

Jana Holsanova

Introduction

In his book *Visual thinking*, Arnheim (1969) writes: “... cognitive operations called thinking are not the privilege of mental processes above and beyond perception but the essential ingredients of perception itself. I am referring to such operations as active exploration, selection, grasping of essentials, simplification, abstraction, analysis and synthesis, completion, correction, comparison, problem solving, as well as combining, separating, putting in context” (Arnheim 1969:13). He draws the conclusion that “visual perception (...) is not a passive recording of stimulus material but an active concern of the mind” (Arnheim 1969:37). In order to understand the dynamics of the ongoing perception and the cognitive processes involved, I investigate the dynamics of picture viewing and picture description (Holsanova 2001). I connect two disciplines, linguistics and vision research, and combine two methods, spoken language descriptions and eye movement protocols, to cast light on the underlying mental processes.

In the first section of this chapter, I present the data, method and analytical tools of the experimental studies. In the second section, I discuss the results of the empirical studies within the functionalist and information-processing framework. The third section addresses how the observers verbalize their impressions from the picture and how their picture descriptions differ in style. In the fourth section, I consider how the results from the studies contribute to an understanding of the dynamics of the ongoing perception processes – an area on the intersection of artistic and cognitive theories. Finally, in Section 5, I present an eye tracking study on picture viewing and mental imagery. The results are summarized and discussed in Section 6.
Picture viewing and picture description: Data, method and analytic tools in the experimental studies

I am going to present results of empirical studies where informants under different conditions (on-line or from memory) and in different settings (monological or interactive) have described a complex picture. In these studies, cognitively oriented research is used with the expectation that the analysis of eye movement data and spoken language description data can give us hints about the dynamics of the underlying cognitive processes. The trouble is, of course, that we cannot directly uncover the content of our mind. If we want to learn about it, we have to do that indirectly, via spoken language in action (via language's overt manifestations) and via the process of visual focusing (via fixations as overt visual manifestations). The closest we can come to throwing light on the mind is via the verbal and visual foci (see Figure 1).

On the one hand, eye movements reflect human thought processes. It is easy to determine which elements attract the observer's eye, in what order and how often. In short, eye movements offer us a window on the mind. On the other hand, verbal foci formulated during picture description are the linguistic expressions of a conscious focus of attention. With the help of a segmented transcript, we can learn what is in the verbal focus at a given time. In short, spoken language description offers us another window on the mind. Both kinds of data will be used as an indirect source to gain insights about the underlying cognitive processes and about human information processing.

In the first eye tracking study, informants were asked to verbally describe a complex picture which they were simultaneously watching (see Figure 2). This picture contains both states and events, allows different ways and scan paths and thereby also different 'readings'. The process of the visual discovery of the picture was registered using an eye tracker. The spoken language description was

Figure 1. Verbal and visual focus.
Figure 2. The complex picture. The motif is from Nordqvist (1990).

recorded, transcribed and segmented into verbal foci and superfoci (Chafe 1980; Holsanova 2001).

What we attend to during the visual perception and the spoken language description can be conceived of with the help of a spotlight metaphor which provides intuitively a notion of limitation and focus: The picture elements fixated visually lie in the focus of a spotlight and are embedded in a context. The spotlight moves to the next area that pops up from the periphery and will lie in the focus for a while. If we focus our concentration and eye movements on a point, we mostly also divert our attention to that point. By using a visual fixation analysis, we can thus follow along the path of attention deployed by the observer. Concerning spoken language description, it has been shown that we focus on one idea a time (Chafe 1980, 1994). The picture elements described lie in the focus of an attentional spotlight and are embedded in the discourse context.

The aim of the study was to conduct a qualitative sequential analysis of the temporal and semantic relations between clusters of the visual and the verbal data. To be able to do that, I created a temporally ordered multimodal score sheets (for details see Holsanova 2001:99f.). The score sheet in Figure 3 shows the visual behaviour (fixated objects as boxes on line 2) and verbal behaviour (verbal idea units on line 3), synchronised over time. Simple bars mark the borders of verbal foci (expressing the conscious focus of attention), double bars mark the borders of verbal superfoci (thematic clusters of foci that form more complex units of thought). This score sheet makes it possible to analyse what is happening during preceding, simultaneous and following fixations when a larger idea is developed and formulated.

With the help of this new analytic format, I could examine what was lying in the visual and verbal attentional spotlight at a particular moment: Configurations of verbal and visual clusters could be extracted and contents in the focused verbal idea flow and the visual fixation clusters could be compared. So for example, instead of analysing the result of the viewing of ‘three birds in the tree’ in form of
a fixation pattern (Figure 4) and the result of the picture description in form of a transcribed extract (see example 1), I was able to visualise and analyse the process of picture viewing and picture description on a time-coded score sheet (as seen in Figure 5). The multimodal method was applied in a recent project on perception and production in on-line writing (Andersson et al. 2006).

The fixation pattern shows us the path of picture discovery, that is, what objects and areas were fixated by the viewer and for how long. This is, though, a static pattern since it does not exactly visualise when and in what order they were fixated. The circles indicate only the position and duration of the fixations, the diameter of each fixation being proportional to its duration. The lines connecting fixations represent saccades. The white circle in the lower right corner is a reference point: it represents the size of a one-second fixation. Let us now look at the transcript.
Example 1. Extract from the transcript, translated from Swedish: ‘Three birds in the tree’:

| 0310 | (2s) and in the **middle** of the field there is a tree |
| 0311 | . with one (1s) with . three birds’ |
| 0312 | . that are doing different things, |
| 0313 | (1s) one bird is sitting on its **eggs’** in a nest’ |
| 0314 | (1s) and the other bird (LAUGHING) is **singing’** |
| 0315 | at the same time as the . **third . like female bird’** |
| 0316 | (1s) is beating the **rug** or something, |

The transcript of the spoken language descriptions is detailed. It includes verbal features, prosodic features (such as intonation, rhythm, tempo, pauses, stress, voice quality, loudness), and non-verbal features (such as laughter). It also contains hesitations, interruptions, restarts false starts and other features that are typical of speech and that give us additional information about the speaker and the situational context. Each numbered line represents a new verbal focus expressing the content of active consciousness. Verbal focus is usually a phrase or a short clause, delimited by prosodic and acoustic features: it has one primary accent, a coherent intonation contour, and is usually preceded by a pause or hesitation (Holsanova 2001: 15f.). It implies that “one new idea is formulated at a time” (Chafe 1994: 108f.) and “active information is replaced by other, partially different information at approximately two second intervals” (Chafe 1987: 22). Several verbal foci are clustered into superfoci (for example summarizing superfocus or a list of items in the above example, delimited by lines). Verbal superfocus is a coherent chunk of speech, typically a longer sentence, that consists of several foci connected by the same thematic aspect and has a sentence final prosodic pattern. Superfoci can be conceived of as thresholds into a new complex units of thought.

As mentioned previously, in order to follow the dynamics of picture viewing and picture description and to analyse the temporal and semantic relations in this process, I had to synchronise the two data outcomes in time. Let us finally have a look at the new format in Figure 5.

In Figure 5, we are ‘zooming in’ on what is happening in the visual and in the verbal stream: which objects or areas are fixated visually and which areas are focused on verbally. As a result of the analysis of the temporal and semantic relations, I could extract several schematic configurations or patterns — both within a focus and within a superfocus (see Holsanova 2001:102f.). Some of them can be seen above: for example an n-to-one configuration between the visual and the verbal part when the viewer is introducing the tree and the birds, or a series of delays between the visual and the verbal part, or a one-to-n configuration when the viewer is, later on, involved in a detailed description of the three different birds.
Figure 5. Schematic configuration on the multimodal time-coded score sheet: 'Three birds in the tree'.

This section gave an introduction to the method and analytic tools used in a series of empirical studies. It was claimed that by combining spoken language description and eye movement protocols, we can get insights into what the observers found interesting, what drew their attention, and how the scene was perceived. The analytical tool – the multimodal time coded score sheet – enables us to synchronize visual and verbal behavior over time, to follow and compare the content of the attentional spotlight and to extract clusters in the visual and verbal flow. It is a processual visualization that gives us insight into the dynamics of picture viewing and picture description. The method developed is important for revealing principles involved in information processing (see Section 2). In addition, this method allows us to follow the process of bridging the gap between visual information gathering and speech production. In other words, we can examine, how the observers successively verbalize their thoughts about the picture (Section 3). Finally, the combination of the data method illuminates mental processes and attitudes and can thus be used as a sensitive evaluative tool for understanding of the dynamics of the ongoing perception process (Section 4).
Information processing: Temporal and semantic correlations between visual and verbal patterns

By now, the reader is acquainted with the method and the analytic tools, and we can review the results of the analysis. What we will be looking at in the following is the relation between the content of the visual focus of attention (specifically clusters of visual fixations) and the content of the verbal focus of attention (specifically clusters of verbal foci).

Three groups of questions are central: Can we identify comparable units in visual perception and in discourse production? Does the order of units in the verbal description reflect the general order in which information was acquired visually? Is the content of the units in picture viewing and picture description similar?

I was inspired by Wallace Chafe who claims that “... similar principles are involved in the way information is acquired from the environment (for example, through eye movements), in the way it is scanned by consciousness during recall, and in the way it is verbalised. All three processes may be guided by a single executive mechanism which determines what is focused on, for how long, and in what sequence” (Chafe 1980: 16). His thesis implies temporal and semantic correlation between the verbal and the visual data stream. In order to test it, I compared patterns during the visual scanning of the picture and during the verbal description of it. The aim was to find out what is lying in the visual and verbal attentional spotlight at a particular moment.

Before summing up the results, let us have a look at Figure 6. showing the process of picture discovery during the verbal formulation “... I see a tree in the middle ... and four men working around it”. The superfocus (between the double bars) consists of two foci. Note that some objects are partly fixated during pauses in the spoken description (TREE) and some objects described in the second focus are partly fixated under the preceding focus (PETTSON 1, PETTSON 3). The consequence of this intertwined structure is that (a) there is no constant temporal relation between the visual and the verbal foci because of the partial lack of the verbal signal, (b) there is a preparatory visual fixation on Petsson 1 and Petsson 3 during the ‘tree’ – sequence. If we, however, look at the superfocus as a whole, we find agreement.

This example illustrates several principles that become revealed in the analysis of temporal and semantic relations in the focused verbal and visual flow: Visual foci track thought which is then verbalised selectively. To sum up, the results show that there is seldom a 1:1 relation between verbal and visual foci and a perfect temporal and semantic match or overlap between the visual and the verbal stream is very rare. The visual focus is often ahead of speech production. If one visual fixation is connected to one verbal focus, the verbal account is usually lagging behind the visual fixations with 2–3 seconds and this delay-configuration is a part
Figure 6. Complex temporal and semantic clusters: ‘Tree in the middle and four men working around it’.

of a larger unit (we saw the configuration ‘series of delays’ in Figure 4 where the three different birds were listed). Areas and objects are frequently re-examined which results either in multiple visual foci or in both multiple visual and verbal foci. Furthermore, the configurations where multiple visual foci are connected to one verbal focus are often intertwined and well integrated into larger wholes (as we could see in Figure 6). In addition, the results indicate that visual scanning is done both during pauses and simultaneously with the verbal description and thus, occasionally, there is no constant temporal correlation between the streams because of the partial lack of the verbal signal. In other words, if we want to find a clear correlation between spoken language and vision, we have to look at larger units of visual and verbal clusters. It seems that the wider the analytic discourse window, the larger the extent to which similar content can be observed in both data streams. I suggest that the superfocus is a suitable unit of comparison since it represents an entity that delimits separable clusters of both visual and verbal data.

In this section, I summarised and illustrated results of the temporal and semantic correlation between visual and verbal data. The theoretical approach used in the study was closely affiliated with functionalist approach and human information processing. The main result is that if we want to find temporal and semantic correlation between visual and verbal data, we have to look at higher levels of discourse hierarchy, that is, at the superfocus level.

We will come back to the semantic analysis in Section 4.

Variations in picture description: Two different styles

There are a number of interesting questions connected to picture viewing and picture description: how the viewers verbalize visual and spatial information from a non-linguistic source (mostly interesting for information-processing theories
and cognitive linguistics), how the observers bridge the gap between visual information gathering and speech production (mostly interesting for psycholinguists), how viewers structure and formulate their verbal description (mostly interesting for linguists and discourse analysts) and how viewers express their impressions from picture viewing, that is, how they describe their first-person experience from viewing (interesting for philosophers, phenomenologists and artistic sciences).

In this section, I will focus on the variation in picture descriptions since the results can be of interest for many of the above mentioned areas of research. In one of the studies where 12 informants described the picture from memory in an interactive setting for a listener, I found two different styles of description, deploying different perspectives. Attending to spatial relations in the picture was dominant in the static description style, while attending to the flow of time was the dominant pattern in the dynamic style. Consider the following two extracts:

(a) 01 Well it’s a picture
    02 rectangular
    03 and it’s mainly green and yellow
    04 with a blue sky
    05 divided into foreground and background,
    06 in the middle there’s a tree
    uh and in it three birds are sitting,
    07 the lowest bird on the left sits on her eggs
    08 and above her
    09 there is a bigger bird standing up,
(b) 01 It’s quite early in the spring
    02 and Pettson and his cat Findus are going to sow,
    03 Pettson starts with digging up his little garden
    04 then he rakes
    05 and then he sows plants
    06 uh. puts potatoes out later on
    07 and when he’s ready
    08 he starts sowing lettuce,

Extract (a) is a prototypical example of a static description style. In what could be called a technical description, the picture is typically decomposed into fields that are then described systematically, using a variety of terms for spatial relations. In the course of description, informants establish an elaborate set of referential frames that are used for localisations. They give a precise number of picture elements, stating their colour, geometric form and position. Apart from spatial expressions, the typical features of the technical description style are frequent use of nouns, existential constructions (‘there is’, ‘it is’, ‘it was’), auxiliary verbs and passive voice.
In the dynamic style (as in Extract (b)), observers primarily focus on temporal relations and dynamic events in the picture. Although there is no temporal or causal order inherent in the picture, viewers infer it: They explicitly mark that they are talking about steps in a process, about successive phases, about a certain order. The dynamic quality of this style is achieved by a frequent use of temporal verbs, temporal adverbs and motion verbs in active voice. Discourse markers are often used to focus and refocus on the picture elements, and to interconnect them. Apart from the above mentioned features, the informants seem to follow a narrative schema: the descriptions start with an introduction of the main characters, their involvement in various activities, and a description of the scene. Table 1 summarizes the characteristics of the static versus dynamic description style.

We are aware of the fact that non-linguistic aspects such as the informants’ knowledge, interests and experience, may to some extent have influenced the description. Especially the informants’ previous knowledge of the picture, of the story, and of the characters may be the most critical factor for distinction of the two description styles since not knowing the characters may lead to a less dynamic description. But when looking closely at the descriptions, we could exclude this factor since 10 out of 12 informants mentioned the characters by name.

Even the informants’ expertise on painting techniques, farming, fauna and flora could have contributed to the way the picture was perceived and described. In addition, the informants’ associations, their way of remembering things may have played role for the description. For this reason, the instruction and the interactional setting were kept constant and the background and the leisure-time interests have been accounted for in a questionnaire.

This section has been devoted to variations in the picture descriptions. Two different tendencies have been observed: the preference to focus on spatial relations in the picture typical of the static technical description style and the preference to focus on dynamic events and temporal relations in the picture typical of the dynamic narrative style.
Understanding the dynamics of the ongoing perception process

Unlike most of the studies in eye tracking and language production, associated with object identification and labelling on word or sentence level, the study of picture viewing and picture description focuses on higher levels of discourse. The starting point are complex ideas about the picture discovery that are formulated in large units of discourse. Because of this character, the verbal and visual data give us insights into the ongoing perception process. Apart from reporting about WHAT they see, the viewers also focus on HOW the picture appears to them. In other words, the informants are involved in both categorizing and interpreting activities.

In the scan path in Figure 7, we can see the way the picture was discovered and examined during the first minute by the one of the viewers.

In the transcript below we can the beginning of the verbal description performed by the same informant:

<table>
<thead>
<tr>
<th>Transcript</th>
</tr>
</thead>
<tbody>
<tr>
<td>0302</td>
</tr>
<tr>
<td>0303</td>
</tr>
<tr>
<td>0304</td>
</tr>
<tr>
<td>0305</td>
</tr>
<tr>
<td>0306</td>
</tr>
<tr>
<td>0307</td>
</tr>
<tr>
<td>0308</td>
</tr>
<tr>
<td>0309</td>
</tr>
</tbody>
</table>
This piece of description illustrates most of the activities that informants do when describing their impressions from the picture:

- They categorize the referents, states and events associated with the picture content;
- They evaluate the picture elements from a qualitative point of view (describe their colours, sizes and attributes);
- They specify the relations between the picture elements, compare picture elements and use metaphors;
- They consider the compositional aspects and group the picture elements in a certain way;
- They express their attitudes, associations, experience and report about mental states.

As for categorization, the main figure, Pettson, can be described as a person, a man, an old guy, a farmer, or he can be called by his name. His appearance could be described – a weird guy –, his cloths – wearing a hat –, the activity he is involved in can be specified – he is digging – and this activity can be evaluated: he is digging frenetically. Thus, we can envision a description on different levels of specificity and with different degree of creativity and freedom.

The tendencies that could be extracted from the data, are the following: Informants start either with a specific categorisation which was then verbally modified (in a potato field or something; the second bird is screaming or something like that) or, more often, a vague categorisation ‘filler’ is followed by a specification: Pettson has found something he is looking at’ he is looking at the soil’. In other words, a general characteristic of an object is successively exchanged against more specified guesses: ‘he is standing and looking at something’ maybe a stone he has in his hand that he has dug out’. When extracting an activity, for example when formulating ‘Pettson is digging’, the eye movements depict the relationships among picture elements
and mimic it in repetitive rhythmic units, by filling in links between functionally close parts of an objects (face – hand – tool; hand – tool, hand – tool – soil). In terms of cognitive semantics we can say that eye movements fill in the relationship according to a schema for an action.

During evaluations, the observers check the details (form, colour, contours, size) but also match the concrete, extracted features with the expected features of an prototypical/similar object: ‘something that looks like a telephone line or telephone poles that is far too little in relation to the humans’. They compare objects both inside the picture world and outside of it. By using metaphors from other domains they compare the animal world with the human one: ‘one bird is looking very human’; ‘there’s a dragonfly like a double aeroplane’.

Concerning introspection and report on mental states, let us look at the following example where one informant, after one minute of picture viewing and picture description, verbalises the following idea: when I think about it then it seems as if it in fact were two different fields, one can interpret it as if they were in two different fields’ those persons here.

When inspecting the visual cluster (Fig. 8), we can see that the observer is rescanning several earlier identified objects that are distributed in the scene. This is consistent with Yarbus' (1967) finding that eye movements occur in cycles and that observers return to the same picture elements several times. This phenomenon has been confirmed by Noton and Stark (1971) who coined the word ‘scanpath’ to describe the repetitive fixations in scene viewing. What we do not see and know,
However, is how the observer perceives the objects on different occasions. I would claim that the objects that are refixated represent a bigger (compositional) portion of the picture and give support for his reconceptualization. By mentally zooming out, he discovers an inferential boundary between parts of the picture which he has not perceived before. The scene originally perceived in terms of 'one field' has become 'two fields' as the observer gets more and more acquainted with the picture. This example illustrates the dynamics of that the observer's perception of the picture unfolds dynamically and can change over time.

In sum, the combination of visual and verbal data shows that the objects are focused on and conceptualised on different levels of specificity, objects' location and attributes are described and evaluated, judgements about properties and relations between picture elements are formulated, metaphors are used as a means of comparison and finally, the object's activity is described. All this involves interpretation and creativity. What we have witnessed here is a process of stepwise specification, evaluation, interpretation and even reconceptualisation of picture elements and the picture as a whole.

Spatial, semantic and mental grouping

A second point I would like to make in this section is about grouping. According to Gestalt psychology, the organizing principles which enable us to perceive the patterns of stimuli as meaningful wholes are defined as (a) proximity, (b) similarity, (c) closure, (d) continuation, and (e) symmetry. The proximity principle implies that objects placed close to each other appear as groups rather than a random cluster. The similarity principle means that there is a tendency for elements of the same shape or colour to be seen as belonging together. Finally, the symmetry principle means that regions bounded by symmetrical borders tend to be perceived as coherent figures.

We can think of several principles of grouping that could have guided the informants' picture description: spatial proximity principle (objects that are depicted closely to each other are described together), categorical or taxonomical proximity (elements that are similar to each other are described together), compositional principle (units delimited by composition are described together), saliency principle (expected, preferred and important elements are described first), animacy principle (human beings and other animate elements are described first) or even other principles.

Let us start with spatial proximity. When we consider a group of objects in the scene (for example the three birds in the tree), it is claimed that observers tend to fixate the closest objects first. This phenomenon is very coercive (because of the proximity to the fovea) and is described as an important principle in the
literature (Lévy-Schoen 1969, 1974, referred by Findlay 1999). And in fact, informants in my study seemed to be partly guided by this principle: they fixated clusters of objects depicted with spatial proximity. However, observers also fixated and described clusters which were not spatially close.

One type of cluster was based on grouping of multiple similar concrete objects (four cats, four versions of Pettson). This type of cluster could not be based on spatial proximity because the picture elements are distributed over the whole scene. In this case, we could rather speak about categorial proximity. (The simultaneity of the objects involved in different activities has probably promoted this guidance.) This type of cluster is usually described in a summarising superfocus that is typically followed by a list where each instance is described in detail.

Another type of cluster that was perceived as a meaningful unit in the scene was hills on the horizon. The observers’ eyes followed the horizontal line, filling in links between objects. This cluster was probably a compositionally guided cluster. The observer is zooming out, scanning picture elements on a compositional level.

We are now moving further away from the scene inherent spatial proximity and coming to the next type of clusters constructed by the observers. This time, the cluster is an example of an active mental grouping of concrete objects based on an extraction of similar traits and activities (see Figure 9: flying objects). Despite the fact that the objects are distributed across the whole scene, they can be perceived of as a unit because of the identified common denominator. The observer is mentally zooming out and creating a unit relatively ‘independent’ of the ‘suggested’ meaningful units in the scene. The eye movements mimic the describers’ functional grouping of objects.
Figure 10. Fixation patterns by a person saying “It looks like in early summer”.

In a number of cases, especially in later parts of the observation, the clusters were based on thematic aspects and guided by top-down factors. In the following example (Figure 10), the observer is verbalising his impression about the picture content: “it looks like an early summer”. Previous scanning of the scene has led the observer to an indirect conclusion about the season of the year. In the visual fixation pattern, we can see large saccades across the whole picture composition. It is obviously a top-down guided cluster, based on mental coupling of concrete objects, their parts or attributes (such as flowers, leaves, colours) on a higher level of abstraction.

Since the concept “early summer” is not directly identical with one or several concrete objects in the scene, the semantic relation cannot be a mere referential one. Instead, the relation is inferred, the objects being concrete indices of a complex (abstract) concept. In addition, the relation between the spoken description and the visual depiction is not a categorical one, associated with object identification. Instead, the observer is formulating how the picture appears to him, on an abstract level. Afterwards, during visual rescanning, the observer is searching again for concrete objects and their parts as crucial indicators for his abstract statement. By refocusing these elements, the observer is in a way collecting evidence for his statement. In other words, he is checking whether the object characteristics in the concrete scene match with the symptoms for the described scenario. Concrete objects can be viewed differently on different occasions as a result of our mental zooming in and out. We have the ability to look at a single concrete object and
simultaneously zoom out and speak about an abstract concept or about the picture as a whole. When talking about creativity and freedom, this type of ‘mental groupings’ shows a high degree of creativity.

In sum, apart from scene inherent concrete picture elements with spatial proximity, even new creative groupings were described on the way. Objects across the scene – horizontally or vertically aligned – were grouped due to the scene composition. Multiple similar elements distributed in the scene were clustered on the basis of a common taxonomic category. Active mental groupings were created on the basis of similar traits and common activity. The process of mental zooming in and out could be documented, where concrete objects were refixedated and viewed on another specificity level or with another concept in mind. During their successive picture discovery, the informants also created new ‘mental’ groupings across the whole scene, based on abstract concepts.

**Picture viewing and mental imagery**

In this section, we will be considered with mental imagery. Arnheim (1969), together with many others, puts the question whether we think in images and what form mental images possibly have. Finke (1989:2) defines mental imagery as “the mental invention or recreation of an experience that in at least some respects resembles the experience of actually perceiving an object or an event, either in conjunction with, or in the absence of, direct sensory stimulation.” It has been proposed that we use mental imagery when we mentally invent or recreate personal experience, when we read novels, plan future events, retrieve information about the physical properties and relations of objects, imagine transformations by mental rotation and mental animation and when we solve problems. A large number of eye tracking studies have been conducted to test mental imagery.

Brandt and Stark (1997) and Laeng and Teodorescu (2001) have shown that spontaneous eye movements closely reflect the content and spatial relations from the original picture or scene. Holsanova, Hedberg, and Nilsson (1998) found similar tendencies as Brandt and Stark (1997), with the difference that the original picture showed a natural, real life scene. In order to extend the finding of previous studies, we chose Sven Nordqvist’s (1990) picture of high complexity as our visual stimulus and conducted a number of new eye tracking studies (Johansson, Holsanova, & Holmqvist 2005, 2006). In the following, I will mention one of our experiments. Twelve students at the University of Lund, 6 females and 6 males, volunteered to participate in the experiment. They were asked to keep their eyes open so that we could film their pupil. The experiment consisted of two main phases, a viewing phase in which the subjects inspected the stimulus picture and a description phase in which the participants with their own words described this picture
from memory while looking at a white screen. Eye movements were recorded during both phases. At the beginning of the viewing phase, each subject received the following instructions:

"You will see a picture. We want you to study the picture as thoroughly as possible. While you study the picture we will measure your pupil size".

The picture was shown for about 30 seconds, and was then covered by a white screen The following description phase took usually 1–2 minutes. The subjects' descriptions were transcribed in order to analyze when certain picture elements were mentioned. The eye movements were then analyzed according to objects derived from the descriptions. For instance, when a subject formulated a following superfocus,

01:20 – And ehhh to the left in the picture
01:23 – there are large daffodils,
01:26 – it looks like there also sat some animals there perhaps,

we would expect the subject to move her eyes towards the left part of the white screen during the first focus. Then it would be plausible to inspect the referent of the second focus (the daffodil). Finally, we could expect the subject to dwell for some time within the daffodil area – on the white screen – searching for the animals (three birds, in fact) that were sitting there on the stimulus picture.

Eye movements were considered correct in local correspondence when within a certain time interval moving from one position to another in the correct direction. Eye movements were considered correct in global correspondence when moving from one position to another and finishing in a position that is spatially correct relatively the whole eye-tracking pattern of the subject (for a detailed description of our method, cf. Johansson et al. resubmitted). We used the Wilcoxon Signed-Ranks test for significance between the number of correct eye movements and the expected number of correct movements by chance.

As Table 2 shows, our results are significant both in the local and in the global correspondence coding suggesting that subjects visualize the spatial configuration of the scene as a support for their descriptions from memory. The effect we measured is strong. More than half of all picture elements mentioned had correct eye movements, according to the conservative global correspondence criteria. Allow-

<table>
<thead>
<tr>
<th>% objects with correct eye movements</th>
<th>Direction significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local correspondence coding 74.8%</td>
<td>W = 76, z = 2.96</td>
</tr>
<tr>
<td></td>
<td>p = 0.0015</td>
</tr>
<tr>
<td>Global correspondence coding 54.9%</td>
<td>W = 66, z = 2.57</td>
</tr>
<tr>
<td></td>
<td>p = 0.0051</td>
</tr>
</tbody>
</table>
ing for re-centering and resizing of the image – as with local correspondence – makes almost three quarters of all picture elements have correct eye movements. Our data indicate that even for a complex picture, spatial locations are to a high degree preserved when describing it from memory.

The resizing effects, i.e. the fact that subjects shrunk, enlarged and stretched the image, were quite common during picture description. It was also common that subjects re-centered the image from time to time; thus yielding local correspondence. Overall, there was a good similarity between data from the viewing and the description phases, as can be seen in Figure 11.

The subjects’ spatial pattern of eye movements was highly consistent with the original spatial arrangement. According to Kosslyn (1994), distance, location and orientation of the internal image can be represented in the visual buffer and it is possible to shift attention to certain parts or aspects of it. Laeng and Teodorescu (2001) interpret their results as a confirmation that eye movements play a functional role during image generation. Mast and Kosslyn (2002) propose, in a similar way as Hebb (1968), that eye movements are stored as spatial indexes that are used to arrange the parts of the image correctly. Our results can be interpreted as further evidence that eye movements play a functional role in visual mental imagery and that eye movements indeed are stored as spatial indexes that are used to arrange the different parts correctly when a mental image is generated.

There are, however, alternative interpretations. Researchers within the ‘embodied’ view claim that instead of relying on an internal image, we use features in the external environment. An imagined scene can then be projected over those external features and any storing of the whole scene internally would thus be unnecessary. Ballard et al. (1997) suggests that subjects leave behind ‘deictic pointers’ to locations of the scene in the environment, which later may be perceptually accessed when they are needed. Pylyshyn (2001) has developed a somewhat similar approach to support propositional representations and speaks about ‘visual indexes’. Another alternative account is the ‘perceptual activity theory’ suggesting that instead of storing images, we store a continually updated and refined set of procedures or schemas that specify how to direct our attention in different situa-
tions (Thomas 1999). In this view, a perceptual experience consists of an ongoing, schema-guided perceptual exploration of the environment. Imagery is then the re-enactment of the specific exploratory perceptual behavior that would be appropriate for exploring the imagined object as if it were actually present. A somewhat similar approach is favored by Barsalou (1999).

In this section, I introduced a study in picture viewing and mental imagery. A significant similarity was found between (a) the eye movement patterns during picture viewing and (b) those produced during picture description (when the subjects were looking at a white screen). The eye movements closely reflected the content and the spatial relations of the original picture suggesting that the subjects created some sort of mental image as an aid for their descriptions from memory.

Summary and discussion

Current theories in visual perception stress the cognitive basis of art and scene perception. We “think” art as much as we “see” art (Solso 1994). Our way of perceiving objects in a scene can be triggered by our expectations, our interests, intentions, previous knowledge, context or instructions.

The aim of the presented studies was to show how overt verbal and visual protocols can, in concert, elucidate covert mental processes. This combination of data illuminates mental processes and attitudes and can thus be used as a sensitive tool for understanding of the dynamics of the ongoing perception process. The results can in this way also contribute to the area on the intersection of artistic and cognitive theories. My data suggest that it is not only recognition of objects that matters but also how the picture appears to the viewers. Verbal descriptions include quality of experience, subjective content and mental state. Viewers report about (i) referents, states and events, (ii) colours, sizes and attributes, (iii) compositional aspects, (iv) they mentally group the perceived objects into more abstract entities, (v) compare picture elements, (vi) express attitudes, associations, (vii) report about mental states.

Subjects start by looking at picture-inherent objects, units and gestalts. As their picture viewing progresses, they tend to create mental units that are more independent of the concrete picture elements. They may make large saccades across the whole picture, picking up information from different locations to support concepts which are distributed across the picture (like “early summer” or “flying insects”). With the increasing cognitive involvement, observers and describers tend to return to certain areas, change their perspective and reformulate or recategory the scene. It becomes clear that perception of the picture changes over time. The dynamics of this categorization process is reflected in the usage of many refixations in picture viewing and reformulations, paraphrases and modifications in
picture description. Finally, during description from memory, the subjects seem to "redraw" or "reconstruct" certain features from the original picture with the help of mental imagery.

References


