

Gist and its role in memory for images

Citation for published version (APA):

Montfort, van, X. A. N. D. R. A. (2006). *Gist and its role in memory for images*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Industrial Engineering and Innovation Sciences]. Technische Universiteit Eindhoven.
<https://doi.org/10.6100/IR609662>

DOI:

[10.6100/IR609662](https://doi.org/10.6100/IR609662)

Document status and date:

Published: 01/01/2006

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Gist

**and its role in
memory for images**

Xandra van Montfort

This research was supported by the Netherlands Organisation for Scientific Research (NWO). The work presented in this thesis was carried out under the auspices of the J.F. Schouten School for User-System Interaction Research.

Printing: Universiteits drukkerij Technische Universiteit Eindhoven

An electronic copy of this thesis is available from the site of the Eindhoven University Library in PDF format (<http://www.tue.nl/bib>).

© X.A.N.D.R.A. van Montfort, 2006

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Montfort, Xandra A.N.D.R.A. van

Gist and its role in memory for images / by Xandra Alice Nicole Danielle Renée Aimee van Montfort. - Eindhoven: Technische Universiteit Eindhoven, 2006. – Proefschrift. -

ISBN 90-386-0696-6

ISBN 978-90-386-0696-5

NUR 778

Keywords: Visual cognition / Change blindness / Gist / Difference detection / Visual memory / Subjective interpretation

Gist and its role in memory for images

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de Rector Magnificus, prof.dr.ir. C.J. van Duijn,
voor een commissie aangewezen door het College voor Promoties
in het openbaar te verdedigen
op dinsdag 20 juni 2006 om 16.00 uur

door

Xandra Alice Nicole Danielle Renée Aimee van Montfort

geboren te Peize

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. D.G. Bouwhuis
en
prof.dr. E.O. Postma

Preface

The work described in this thesis was conducted as part of the NWO funded ToKeN research program (ToKeN stands for: *To accessibility and Knowledge Enhancement in the Netherlands*). For the current project, entitled Eidetic, researchers from the Universities of Delft, Maastricht, Nijmegen and Eindhoven jointly confronted the challenge of retrieving visual information by focusing on both the computational and usability aspects of content-based image retrieval (CBIR). Whereas the researchers from the other three universities focused on the possibilities of computerized image retrieval, the research described in this thesis focuses on human performance.

In addition to NWO, there are others who have contributed to the work described in this thesis, scientific and/ or otherwise. I would like to thank the following persons for their contribution: my advisors Don Bouwhuis and Paul de Greef; Frans Verstraten, Jack van Wijk and Ronald Rensink (for useful comments on an earlier version of this thesis); Frans Blommaert and Jan van Bolhuis; the research group "Decision Making" (for letting me join their experimental sessions). I am especially grateful to Iris van Rooij and Eric Postma, who both contributed a lot to the quality of this thesis in a relatively short period of time.

I would like to take this opportunity to express my gratitude to Gideon Keren. From the moment he became my advisor during my masters' thesis he has encouraged me to pursue a life of science by his unfaltering belief in my work and his enthusiasm. He was always available for questions and actively helped me find other researchers who would be able to help when necessary. Thank you.

I would also like to thank Astarte for providing me with a large supply of naive participants for pilot studies, my parents for providing feedback on text and moral support. I am also immensely grateful to Audrey and WP for their moral support and their willingness to Rate 770 descriptions for 72 images in return for a dinner I would have paid for anyway. Finally I would like to thank Seth, for doing all of the above.

Contents

Preface	v
Contents.....	vi
1 Introduction	1
1.1 Content-based image retrieval.....	2
1.2 Human memory for images	3
1.3 The aim of this thesis.....	4
2 Gist: That is what it is all about	7
2.1. Defining the concept gist and characteristics associated with it	7
2.2 Context-dependency of image interpretation and gist descriptions	10
Discussion and conclusion	13
3 A method to determine a gist change.....	15
3.1 How to determine a change in gist	15
3.2 Proposed method to determine a gist change	18
3.3 Validity of the method.....	22
Conclusion and Implications	23
4 Validation of test stimuli.....	25
4.1 Considerations for valid test images	25
4.2 Determining validity of the changes	33
Conclusion.....	41
5 Measuring change detection	43
5.1 Evidence for importance of semantic meaning	44
5.2 Experiment 1: Intentional encoding	46
5.3 Experiment 2: Incidental encoding.....	51
5.4 Comparing results from Experiment 1 and Experiment 2.....	55
General discussion and conclusions	56
6 More than you compare.....	57
6.1 Previous evidence for preserved representations	58
6.2 Experiment 3: a Same/Different task and a 4AFC task.....	61

Conclusion.....	68
7 Conclusion.....	69
7.1 Discussion and implications.....	69
7.2 Preliminary results from the Flicker paradigm.....	72
7.3 Conclusion.....	75
References	77
Appendix	83
Summary	85
Samenvatting	88
Biography	91

1

Introduction

A picture is worth a thousand words. This expression reflects the common experience that an image can convey an idea more effectively than verbal communication. However, the improved effectiveness comes at a prize.

Take a moment to look at the image in Figure 1.1. Will you recognize the picture if you see it again sometime next year? What properties of the picture will you still remember one year from now? The research in this thesis will demonstrate that you will remember the essence of the image: the "gist".



Figure 1.1.

Although questions such as stated above might seem artificial at first, answers to such questions in general are relevant for numerous research areas. For instance, answers can provide insight in human visual memory, what kinds of changes people will notice, how

reliable eye-witness testimonies are (e.g., Loftus & Ketcham, 1991), and what people will probably remember from television commercials. In this chapter the relevance of a better understanding of human memory for images is clarified by considering the processes enabling someone to find an image in a large digital image database. Section 1.1 shortly discusses the current state of computerized image retrieval systems. Given the current facilities of computerized image retrieval, it is very likely that part of successful search for one particular image depends on what people remember from an image. Section 1.2 discusses the limitations of human accuracy concerning visual recognition and visual memory. Section 1.3 states the aim of this thesis and the research questions that will be addressed in the research reported in this thesis. This chapter ends with an outline of the following chapters.

1.1 Content-based image retrieval

The fact that an image can convey an idea more effectively than words, might have contributed to the increasing popularity of digital image databases. As more and more digital image databases become publicly available (e.g., museum collections), the need for user-friendly image retrieval systems increases. The classical image retrieval systems are text based and retrieve images by looking for specified key words. Such a retrieval system requires images to be accompanied by labels facilitating retrieval. These labels have to be assigned. As the amount of available images is growing rapidly the task of labeling all images is enormous. Not only is the number of images very large, also the number of possible key words for each image is very large, as "an image is worth a thousand words".

One way to sidestep the need for manually labeling all images in a large database is by using Content-Based Image Retrieval (CBIR). CBIR strives to enable computers to extract information directly from an image. (Smeulders, Worring, Santini, Gupta and Jain (2000), provide a good review). When computers would be able to detect and recognize objects, it would be possible to let the computer retrieve all images with a particular object specified by a user. Consider the processes needed to recognize partly occluded objects to appreciate the difficulty associated with computerized object recognition. The current state of progress on the topic of computer vision is that computers are pretty good at extracting information at a pixel level (low-level). Examples of recent research on CBIR techniques using low-level information are automatic detection of watermarks and chain lines in paper, for the purpose of authenticating etchings (van Staalduinen, van der Lubbe and Backer, 2004), and image retrieval based on color, texture and shape or a combination of the three features as thoroughly described by van den Broek (2005). At a (somewhat) higher level of perception, Bergboer, Postma and van den Herik (in press) have devised a method, which enables computers to detect faces by making use of context information. Although with the present available content-based image retrieval techniques it is possible to reduce the amount of images that might be relevant by applying a search query, in most cases it will result in numerous possible targets.

When considering the task of retrieving a previously seen image with help of a CBIR search engine, the human (user) task is twofold. First, the user has to define the search task. Second, the user has to recognize the target image. For both tasks the user will have to rely on what he or she has remembered from the image. Surprisingly little is known of exactly which properties people remember from an image and subsequently use for search and recognition.

1.2 Human memory for images

In general people are very good in visually recognizing things (e.g., scenes, images, faces) they have seen before. Shepard (1967) and Standing (1973) demonstrated the huge capacity for remembering and recognizing pictures under experimental conditions more than thirty years ago. Initially the apparent large capacity for memory of pictures was interpreted as evidence that human visual long-term memory is virtually unlimited. However, numerous experiments have demonstrated that humans are not particularly good at precise recognition. In this section three examples of flaws in human visual memory are presented. Then, the implications, of these flaws, for human visual memory are discussed.

Limits to human visual memory

Although everyday visual experience appears rich and detailed, research has demonstrated that visual memory might not be as rich and detailed as it seems to be. For instance, when in a test phase some previously studied images are mirrored, human subjects are not very accurate in reporting whether the orientation was the same (e.g., Intraub, 1980, Standing, Conezio, and Haber, 1970). Also, Nickerson and Adams (1979) demonstrated that 50% of the American subjects failed to correctly identify a penny, an item they encountered almost daily for years, from a set of visually similar distractors. When an object from a studied image is replaced by another object, changed in size, changed in position, inserted or deleted, participants do not always notice the difference (Mandler and Johnson, 1976, Friedman 1979, Mondy and Coltheart, 2000). These results were interpreted as evidence that visual details are not retained in memory (Friedman, 1979).

People are also insensitive to changes applied during a visual disruption (termed Change Blindness). Large changes in images, movie fragments as well as in real life situations can go undetected (see Simons and Levin, 1997 for an overview). Perhaps the most striking and counterintuitive demonstration of Change Blindness is a study in which 50% of the participants failed to notice that the person they were talking to was substituted by another person during a brief visual disruption (Simons and Levin, 1998). See Rensink (2002) for a detailed overview of research on the topic of change detection.

In contrast to studies suggesting limited memory for visual information, other studies demonstrate that people remembered *more* from their visual surroundings than was actually visible. The phenomenon "boundary extension" is one example of circumstances in which people seem to remember more than they have actually seen (Intraub and Richardson, 1989). When participants were asked to draw what they remembered from a picture they had studied previously, the drawings tended to contain more information than was present in the studied picture. For instance, for a picture displaying a collection of garbage cans with some cans partially displayed, in the redrawn versions of the picture, the garbage cans would be completely drawn. Another example of circumstances in which participants recall more than was actually present is when the recalled objects are expected to be present. For instance, people falsely reported having seen books in an office, whereas the office did not contain any books at all (Brewer & Treyans, 1981). The question that arises is: what and how much visual information do people encode in memory?

Implications of limits for visual memory

To account for the demonstrated difficulty in detecting changes, it was suggested that human visual memory is schematic and abstract. People encode the so-called gist of a scene/image and details are forgotten or overwritten (Intraub, 1997; Friedman, 1979; McConkie & Currie, 1996; Rensink, O'Regan, & Clark, 1997; Simons, 1996; for a review see Rensink, 2000(a) and Simons & Levin, 1997). In fact, Friedman (1979) suggested that changes not affecting the gist of an image are unlikely to be detected (see also Simons & Levin, 1997).

Recent evidence however suggests that human visual memory might be more detailed than change detection theories suggest. According to this evidence a relatively detailed scene representation is built up in memory across eye fixations (Henderson and Hollingsworth, 2003; Hollingsworth, 2005; Melcher, 2006). Change Blindness is explained by either a lack of attending the changed aspect during the initial viewing of the image, or a lack of attending and comparing the image with the stored information during the test phase (Hollingsworth & Henderson, 2002). This would suggest that failed change detection does not necessarily imply limited memory, but might be caused by a failure of comparison between the stimulus and the stored information (see Simons and Rensink, 2005, for an overview). This explanation of Change Blindness differs from the traditional one in predictions for the fate of visual details in memory. According to the traditional view visual detail information is overwritten or forgotten. In the new view detailed visual information is retained. Although the extent and nature of visual memory is currently still debated in the field of visual cognition, there has been consensus for years that people are likely to remember the essence of an image; the gist. The question remains what exactly is the gist of an image, and what is its role in memory for image?

1.3 The aim of this thesis

Although it is commonly assumed that the gist of an image is remembered, and changes affecting the gist will be detected, this has not been tested directly. One possible reason for the lack of direct evidence is that in the domain of visual cognition, the term gist is poorly defined (Henderson & Ferreira, 2004). There is no agreement on which visual properties carry the gist of an image (Wolfe, 1998). There is consensus that the term gist refers to the high-level meaning or the essence of an image. Intuitively, people might have a clear understanding of what is the gist of an image. Clearly, the gist of an image is an interpretation of the image. At a physical level an image is just a collection of pixels arranged in a particular way (the physical features). When humans perceive an image, they interpret it by relying on their experience and knowledge of the world to recognize the objects of the image (the object features). The meaning of an image (the semantics) is subjective, and can therefore vary between different observers and vary with context. For instance, the image in Figure 1.1 might be interpreted as an image of a Dracula or a Vampire. It can also be interpreted as an image of a woman being attacked, or more specifically: a blond woman being attacked by a dark haired Dracula. The interpretation of an image can be context dependent. So an image can have a different meaning for different people and in different situations. To test the role of gist for memory of images, a firm understanding of the concept of gist is essential. Then the role of gist for memory for images can be tested by examining whether people are able to detect a difference between a studied image and an altered version of the image in a test

phase. By testing different types of alterations (alterations affecting the gist of an image and alterations that do not affect the gist) the role of gist for memory for images can be examined.

As there is no consensus on the properties that are part of the gist, it is even more problematic to define changes affecting the gist. To be able to assess the role of gist for memory for images, three research questions are addressed in this thesis. The first question addressed is:

1. What is the gist of an image, and what are changes affecting the gist?

With a firm understanding of the concept of gist, and a method to determine whether or not a change affects the gist, the second research question is addressed:

2. Are changes affecting the gist better detected than other changes?

Recently it has been suggested that failing change detection is not necessarily a result of limited memory. This implies that a better detection of changes affecting the gist of an image might be caused by some other factor apart from better memory. Whether or not the gist of an image is the only visual information people remember from an image is addressed by the last research question.

3. Is the gist of an image used more often than other image properties during a change detection task?

Answers to the three research questions will provide an understanding for what people are likely to remember from an image.

Outline

In chapter 2 several definitions of gist and the assumed as well as the demonstrated characteristics of gist, as described in the literature, are reviewed. This review reveals that the subjective nature of gist is often ignored in literature referring to the concept of gist. An operational definition of gist is proposed that incorporates the subjective nature of interpreting images. Some expectations regarding variations in image interpretation are discussed. The notion of gist as a subjective interpretation of an image, as defined in chapter 2, leads us to identify the challenge of determining *the* gist of an image.

In chapter 3 a method is proposed to sidestep the need to know what *the* gist of an image is, by detecting a *change* in gist between two images. The method was developed to test whether two images (one derived from the other) are systematically interpreted equal or different. The rationale behind the method is explained by considering an example with simple stimuli. In the last part of chapter 3 the validity of the proposed method is addressed.

To examine whether changes to an image affecting the gist of the image are better detected than other changes, a special set of test stimuli was developed. In chapter 4 some other possible factors which could influence the visibility of changes are discussed. Further, the set of test stimuli is validated to ensure that a change affecting the gist is indeed perceived by participants as a change affecting the gist. In addition, it is verified that a change not supposed to affect the gist indeed does not change the gist. To this purpose, the method proposed in chapter 3 is applied to the developed test stimuli.

Chapter 5 sets out to test whether changes affecting the gist of an image are better detected than other changes. First, evidence from other studies addressing the sensitivity for changes to meaningful (semantic) properties is discussed. Then, the sensitivity for gist changes is tested in two experiments. In the first experiment participants are informed of an upcoming memory task; in the second experiment they are not.

Chapter 6 tests whether the gist of an image is better retained in memory than other visual information. Earlier studies are reviewed which provide evidence that a change detection task does not necessarily accurately reflect how much information participants remembered from an image. It is examined to what extent this pertains to gist and other visual information. In the reported experiment recognition performance and difference detection performance are compared.

In chapter 7 the implications and limitations of the reported research are addressed. Finally, conclusions are drawn to provide an answer to the research questions stated in this chapter (section 1.3).

2

Gist: That is what it is all about

The research discussed in this thesis aims at providing a better understanding of the concept of gist and its role in memory for images. The central concept in addressing this is gist. In the Oxford American Dictionary of Current English (1999) gist is defined as: "The substance or essence of a matter". In the field of visual cognition the concept of gist is poorly defined (Wolfe, 1998; Henderson & Ferreira, 2004). In this chapter several definitions of gist are summarized, and some of the characteristics often associated with gist are addressed. Further, one characteristic that is often not explicitly addressed in the literature is discussed: the characteristic that the gist of an image might not be equal for different observers. For the research in this thesis an operational definition of gist is proposed. The operational definition explicitly leaves room for subjective interpretations of an image. This chapter ends with a discussion of some expectations on possible factors affecting the interpretation of an image, and the way people will describe it.

2.1. Defining the concept gist and characteristics associated with it

The main difficulty with investigating the gist of images is that it is poorly defined. Although everybody has an intuitive notion of the gist of an image, the variety of definitions makes it hard to predict what people will remember and which applied changes people might notice.

The gist of a scene is usually referred to as a short description capturing the essence or identity of the scene (e.g., Rensink, 2000b) or its central theme (e.g., Potter, Staub & O'Conner, 2004). The various definitions of gist, found in the field of visual cognition, all suggest that the gist of an image is the high-level semantic meaning of its contents. Such definitions omit specification of the visual properties that carry this high-level meaning. As a result, there is no consensus on the definition/content of gist (Wolfe, 1998). Descriptions of the content of gist vary with the context and/or the researcher employing the concept. For instance, Oliva (2005) suggests that the gist of an image contains spatial information (i.e., layout), whereas Henderson & Hollingsworth state at some point that "... scene memory is limited to gist, layout, and perhaps the abstract identities of recognized objects ..." (Henderson & Hollingsworth, 2002, page 116). This implies that according to Henderson and Hollingsworth layout is not part of the gist.

In difference detection experiments, participants are instructed to detect the difference between two subsequent images. To detect the difference, the relevant properties of an image need to be perceived during the first encounter, and subsequently stored and retained in memory. Upon presentation of the second image, the retained information is compared to the perceived information. Three characteristics of gist that are generally associated with difference detection are: (1) gist perception is fast, (2) gist is stored in memory, and (3) gist is compared. Below these three characteristics are briefly discussed.

Gist perception is fast

The first characteristic often attributed to gist is that it is perceived very quickly. There is abundant evidence for the fast comprehension of the meaning of an image (Potter, 1975, 1976; Oliva, 2005; Oliva & Schyns, 1997; Intraub, 1980, 1981; Biederman, Rabinowitz, Glass, & Stacy, 1974; Biederman, 1981; Thorpe, Fize, Marlot, 1996; Rousselet, Joubert, & Fabre-Thorpe, 2005). An example of research providing evidence for the fast extraction of gist from an image comes from a series of experiments by Potter (1975, 1976). She demonstrated that people quickly comprehend the meaning of an image. Participants saw a number of images in rapid succession (e.g., 125 ms per image). Participants were instructed to respond when they recognized a specified target image. The target image was either specified by a preview of the image or by a short verbal description capturing the meaning of the image (e.g., two men drinking beer, a child and butterfly). Detection performance for previewed and verbally described targets was not significantly different. As correct identification of a verbally described target image requires that the image in the test phase is identified and compared to the verbal description of the target image, Potter concluded that images are understood very quickly (less than 125 ms).

Gist is stored in memory

A second characteristic often attributed to gist is that the gist is what people remember (Simons & Levin, 1997; Intraub, 1984; Wolfe, 1998; Friedman, 1979; Biederman, 1981). According to for instance Intraub (1997) people encode a schematic abstract representation of an image (i.e., the gist) while visual details are forgotten. When the concept of gist is used to explain what people remembered of a scene, the frame theory of Friedman (1979) is often cited. The concept of Friedman's frame theory is closely related to the schemata described by Biederman (1981). According to Friedman, when people look at a scene they will recognize objects in a particular setting and this will activate a conceptual frame from memory. The activated frame contains objects or aspects commonly encountered in similar situations. For instance, with a kitchen frame, commonly encountered objects are a fridge, a stove, a coffee maker etc. According to Friedman (1979) the objects associated with the activated frame will receive less attention than objects not commonly encountered objects, during subsequent inspection of an image. The frame theory predicts that expected objects (consistent with the activated frame) are encoded, stored, and retained in memory less detailed than unexpected objects. See also Hollingworth & Henderson (2002) for a distinction between what they call "consistent" and "inconsistent" objects referring to objects commonly and infrequently encountered with the activated frame, respectively. Furthermore, frame theory predicts that people will remember the frame and the objects or aspects that are not commonly encountered in the frame of that particular scene (i.e., unexpected objects). The presence of expected objects is inferred from stereotype frames. Brewer & Treyans (1981) demonstrated the (false)

memory for expected objects. In their study participants falsely reported having seen books in an office (Brewer & Treynans, 1981).

Closely related to the frame-based explanation, is the assumption of Rensink et al. (1997), that the gist includes a description of the most interesting aspects of an image. The interesting aspects are assumed to attract and guide attention. As the interesting aspects will receive more attention during image perception than uninteresting aspects, interesting aspects are more likely to be remembered than uninteresting aspects.

Gist is compared from view to view

A third characteristic associated with gist is that it is compared in change detection experiments. Simons (2000) discussed possible explanations to account for participants' difficulties in detecting changes made during a visual disruption. These explanations range from no visual information is stored, to no visual information is compared. However, all of these explanations for poor detection of changes assume that the gists of the involved images are compared. Also Friedman (1979) suggested that all changes not affecting the gist are not detected (see also Simons & Levin, 1997).

An additional characteristic: gist is subjective

Having discussed the three generally appreciated characteristics of gist, we introduce an important additional characteristic that is largely ignored in the commonly used definitions of gist. This additional characteristic is that the perceived gist of an image depends on the observer. People interpret a scene based both on what they see (image properties) and what they know (knowledge, experience, interests, expectations, etc.). The identification of the high-level category or identity of an image, the activated frame, and the subjective interesting properties of an image, and therefore the gist, can fluctuate for different people. Often images can belong to more than one category. Consider, for instance, an image of a dormitory room with a desk and a bed. Such an image can be categorized as a bedroom, a study, or a dormitory. Similarly, images can activate different frames. For instance, the dormitory image can activate the frame of a bedroom as well as that of a study. The activation of a frame affects what objects are expected or unexpected. In a study a desk would be expected, whereas it would not be in a bedroom. A refrigerator would be unexpected in both the bedroom and the study, while it is not out of place in a dormitory room concept. Also, what aspects are included in the frame depends on the experience and previous knowledge of an observer. Dutch students rarely share a room with someone, whereas in other countries it might be common to have a roommate. So for Dutch students a picture of a room with two beds and two desks might not be associated with the concept of dormitory. Furthermore, people can be different in the aspects of an image they find interesting, and therefore meaningful.

The definition of the gist of an image should facilitate people to have their own opinion on the essence of an image. For the proposed definition of gist in this thesis the main characteristic often attributed to gist, which is taken into account, is that the gist should reflect what is remembered from an image. Therefore, in this thesis gist is defined as: **a subjective interpretation of what an image is about**. Our definition does not specify the visual properties in an image which carry the gist, as the gist, and therefore the properties carrying the gist depend on the context and the person observing the image. Our definition can be

considered an operational definition as it provides a way to tap into the concept of gist. A description of the gist as perceived by a person is obtained by asking the person to verbalize his or her interpretation of what the image is about. The next section contains expectations for the expected level of the verbal description and possible differences in interpretation of images.

2.2 Context-dependency of image interpretation and gist descriptions

Humans are *Homo significans*: meaning makers (Chandler, 1997). When viewing an image people try to impose meaning, they try to make sense of what they see. They will not register every single pixel in the image, but interpret an image on what it is about, who, what, where, when and why. Even when for instance an image is empty (blank canvas) people will find a way to interpret what they see. When two people look at the exact same image, the meaning of the image need not be the same for the two people. As stated in the previous section, the interpretation of an image is subjective; people can differ in their interpretation of what an image is about, and what the meaningful aspects are, the essence (i.e., the gist). Figure 2.1 illustrates the two main sources that affect image interpretation: image properties and observer characteristics. Several context factors possibly affecting the interpretation of an image are discussed. In what follows, we discuss the level of interpretation and the circumstances influencing the interpretation of images.

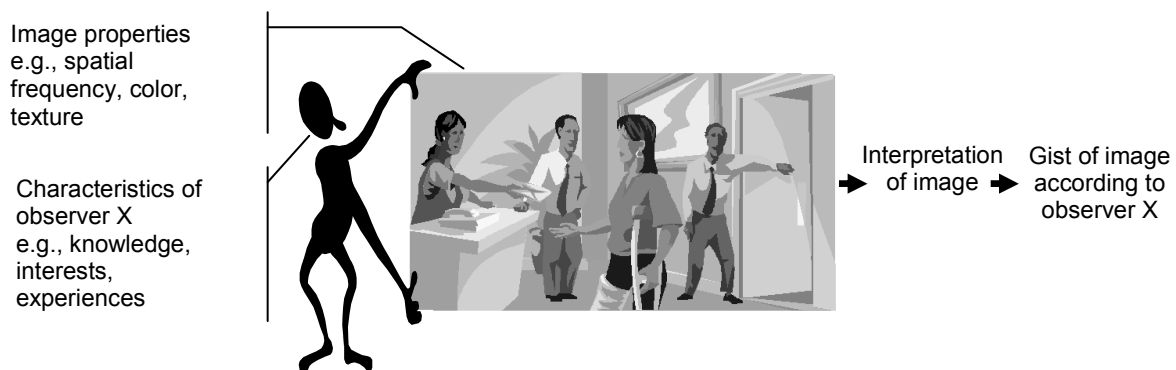


Figure 2.1. The gist of an image depends on image properties and observer characteristics.

Level of identification

An image is at the physical level a collection of colored pixels arranged in a particular way. Images can be interpreted at different levels. The image in Figure 2.1, for example, can be interpreted at a low level (e.g., black, purple, green, yellow, and blue colored pixels), a medium level (e.g., two men and two women, a desk, a telephone, a plant, a door) and a high level (e.g., hospital).

Action Identification Theory (AIT; Vallacher & Wegner, 1987; Wegner & Vallacher 1986) can be used to predict at what level images are interpreted. AIT predicts a preferred level at which people will describe and think about their actions and proposes that these descriptions (identities) can be arranged in a hierarchy. According to AIT there are lower-level and higher-

level descriptions. For actions the lower-level descriptions identify *how* something has to be done whereas higher-level descriptions indicate *what* is done and why. The interpretation of images can follow a similar hierarchical sequence, ranging from perceptual levels, such as color and texture, to conceptual levels for objects and the scene as a whole (Hollink, Schreiber, Wielinga and Worrying, 2004).

The level at which people are likely to identify an action is based on the degree of automatism for the action (Wegner & Vallacher 1986). The more familiar and automatically an action is and can be performed, the higher the level at which it is identified. For instance, when first learning how to walk, children will identify their action as walking. With more experience, the same action will be identified as going somewhere. Similarly, for images the level at which an image is interpreted is based on automatism and knowledge. For instance, when children learn to read, they focus on the individual letters in a word. When more experienced, they will see words, not individual letters. The automatism of word recognition is clearly demonstrated by the Stroop effect (Stroop, 1935). People are faster at identifying a word than at identifying the color in which the letters are printed. Children, who can not read, have much less difficulty with identifying the color of printed words. So, one of the factors determining the level at which an image is interpreted is the degree of familiarity.

Another prediction from AIT for the identification of actions is that people tend to describe their actions at the highest possible level (Wegner & Vallacher 1986). For instance, when facing a particularly difficult mountain descent on a ski trip, the performed action can be identified as "skiing", or as "keeping the speed under control". Although they are both accurate descriptions of the same behavior, keeping the speed under control might be the preferred level of interpretation. Such a preferred level of description can also be found for images. For instance, although most images can be described by noting depicted colors and shapes, this is seldom the preferred level of description. People will try to find meaning (a higher level of description). When no meaning can be found, for instance, in abstract art, an image might be described using low-level properties, such as texture, color or shape. Also, if the highest possible level of identification of an image does not reflect the essence of an image for the observer, a lower-level description is adopted. So when an observer finds that in the image in Figure 2.1, the courtesy of the man at the door is important for the image, the preferred level of description will be at the associated level. When the predictions initially stated for identification of actions are applied to the interpretation and description of images, it is expected that people will describe an image at the highest possible level and will include the meaningful aspects.

A topic related to the preferred level of interpretation of images is the categorization of objects. However, when the concept of gist as proposed in this thesis, and more specifically the preferred level of interpretation of an image, is compared to object categorization theory (see for instance, Rosch, Mevis, Gray, Johnson, & Boyes-Braem, 1976), there are two important differences. Object categorization theory distinguishes between three levels at which an object can be categorized; the superordinate level (e.g., vehicle), the basic or entry level (e.g., car), and the subordinate level (e.g., Ferrari). In object categorization theory the basic level has a special status. Research has revealed that the basic level is the level most frequently used in language, is the level which is first learned and used by children, and is the level at which objects are recognized fastest. The preferred level of interpretation of an image does not have such a special status. In contrast to the basic level of an object, the preferred

level of interpretation does not necessarily reflect what is recognized the fastest. In this thesis the preferred level of interpretation is defined to reflect what people remember. As a consequence, for the preferred level of interpretation objects can be categorized at a more specific subordinate level. For instance, it is possible at first glance, an object is recognized as a car, and only later is recognized as a Ferrari. When the fact that the car is a Ferrari is important for the observer, it is expected that the observer will remember the "type" of car and the type will be reflected in the preferred level of interpretation. Another difference between the concept of basic level of object categorization and the preferred level of interpretation is the fact that the preferred level of interpretation is not necessarily the level used by most people. Whereas the basic level of categorization is more or less stable for people from similar cultures, the preferred level of interpretation is far less stable. The preferred level of interpretation can be different for different people and in different contexts.

Differences in interpretation

In addition to differences in the level at which people interpret an image, people can differ in their interpretation of what an image is about. As described before, images can belong to more than one category. Also, which frame is associated with an image can be different for different people. This has consequences for the idea which properties or objects are unexpected in a scene and what is to be perceived as being inconsistent with a scene. Besides personal differences in interpretation based on the categorization of an image, people can differ in what aspects they find interesting. Someone might describe the image in Figure 2.1 as: "a man opening a door for a woman". Other people may think the woman with the cast receiving an envelope the most meaningful part of the image.

Another example of a context factor possibly affecting the interpretation of an image is time. One image can be interpreted in one way at a particular moment and in another way at another moment. For instance, a picture of the skyline of New York taken in 1999 could have been interpreted as a picture of New York. That same picture might now be interpreted as a picture of the Twin Towers. Knowledge and context can lead to different interpretations of the same image by one person at different times or in a different context.

As the interpretation of an image depends on the observer, there is not one *true* interpretation. It is even possible that different people have seemingly contradictory interpretations. Consider for instance an image of the American army marching into Iraq. An American might interpret the image as bringing freedom, while an Iraqi might interpret the same image as becoming suppressed.

Although people can differ in their interpretation of an image, in general people will be able to see that different interpretations are possible. It is not difficult to imagine that the image in Figure 2.1 can be interpreted as both "a hospital" and "a reception". So even if different people would describe the same image in many different ways, people are capable of judging if a description given by another person is reasonable.

Differences in description

Differences in the preferred level of identification and the interpretation of an image will lead to different verbal descriptions of the same image. But even when people will interpret an image exactly the same and use the same level of identification, they might use different

words to describe their interpretation. This again will lead to differences among descriptions of the same image by different people.

Discussion and conclusion

Where the gist as used by other researchers is often used to signal what is the first that people see in an image, the main objective of the gist as defined in this thesis is to reflect what people are likely to remember from an image. Although no research has been done on exactly how the two concepts differ, it is expected that the gist as defined in this thesis is more elaborate than the traditional concept of gist, as used by researchers such as Potter and Oliva.

One important difference with the concept of gist as used by specifically Oliva (2005) is the fact that the gist as described in this thesis is not necessarily reflected in the image itself. Oliva suggests that the gist of an image can be extracted from the physical properties of the image itself, whereas the gist as defined in thesis mainly relies on the interpretation of the image. In this respect, our gist can be seen to reflect the "Conceptual gist", whereas Oliva distinguishes between a "Perceptual" and a "Conceptual" component. From literature alone it is not clear how the two component of gist as defined by Oliva relate to each other. That is, it is not clear whether for Oliva's conceptual gist a corresponding perceptual gist is an absolute requirement.

The Frame theory (Friedman, 1979) is the concept most similar to the here defined gist. The Frame theory tries, as the gist in this thesis, to define which properties of an image people will remember. Another similarity is that Friedman recognizes the fact that when an observer has a special interest for a certain property (for instance for a stove in a kitchen scene), the property is more likely to be remembered. This prediction is closely related to the subjective nature of gist as stipulated in this thesis.

In sum, the gist of an image is not simply an inventory of what is displayed. In this thesis the gist of an image is defined as a personal interpretation of what an image is about. A description of the interpretation of the essence of an image is assumed to reflect the gist. The level at which an image is interpreted, the categorization of the image, the identification of meaningful aspect and the formulation of a description depends on observer characteristics such as knowledge, interest, and the context. Since gist is an interpretation of the meaning of an image, it is impossible to determine *the* gist of a scene. In our investigation of the role of gist in memory for images by measuring difference detection performance (described in chapters 5 and 6), the question that needs to be answered is not necessarily how to define the gist, but how to measure a *change* in gist. A method to determine if a change affects the gist of an image is proposed in the next chapter.

3

A method to determine a gist change¹

In chapter 2 some commonly used definitions of gist and four characteristics of gist were discussed. The gist of an image was defined as a personal interpretation of what an image is about. In this chapter difficulties in determining a change in gist as a result of a change in an image are addressed. A method is proposed to determine whether or not two images are interpreted differently, and can therefore be characterized as having a different gist. The method has the benefit of taking any subjective interpretations of the researcher, and the subjective interpretations of participants interpreting the images out of the process and turning qualitative information into quantitative data that can be analyzed². This chapter ends with a discussion on the validity of the proposed method.

3.1 How to determine a change in gist

The identification of the essence of a scene, the activated schema or frame, and the interesting properties of a scene, and therefore the gist, can differ for different people. Knowing the exact interpretation of a particular image by different people is not relevant for research on the effects of gist on change detection. For instance, different people can interpret Figure 3.1(b) as either a vampire, Dracula or a man dressed up for Halloween. Although different people may have different interpretations of Figure 3.1(a) and (b), most people will agree that Figures 3.1(a) and (b) have a different gist. When investigating the effects of gist on change detection, the question that needs to be answered is not "how should we define gist?" but instead "how can we determine a *change* in gist?"

As gist is a personal interpretation of what an image is about, a method to measure if a change affects the gist must be based on differences in interpretation within a single person. The gist is affected by a change when people *systematically* interpret the original image differently than the changed version of the image. The images in Figure 3.1 for instance are likely to be interpreted differently as a result of the added fangs in the second image. It is however not always obvious whether two images are interpreted differently. Showing someone the two

¹ This chapter is based on Van Montfort, X.A.N.D.R.A., (accepted for publication). What makes a difference: A method to determine if a change in an image affects the perceived gist. *Behavior Research Methods*.

² Formulation inspired by an anonymous reviewer.

images simultaneously and asking whether they interpret them differently is not sufficient. When a person sees both images side by side, the difference between the two images can become salient; that is, an area not receiving attention when an image is presented alone might receive attention when two images are presented together. For instance, people might only notice the fangs in Figure 3.1(b) because they are searching for a difference between the two images. If the fangs are not noticed when Figure 3.1(b) is presented alone, the image is unlikely to be interpreted differently from Figure 3.1(a), as when the fangs are not noticed, Figures 3.1(a) and (b) are both likely to be interpreted as an image portraying a loving couple. Simply asking people whether a change affects the interpretation may lead to overestimating the relevance of the change for the interpretation of the image. Therefore, showing someone the two images simultaneously and asking whether they interpret them differently is not sufficient to determine whether a change affects the gist of an image or not.



Figure 3.1. Example of a change affecting the gist of the image. The added fangs result in different interpretations of images (a) and (b).

Asking a person to interpret both images and then comparing the two interpretations is also not sufficient to determine a difference in interpretation between two images. When a person is asked to interpret two images, the interpretation of the second image is influenced by the first image (Garner, 1974). This again may lead to overestimating the importance of the difference.

It is better to ask a person to interpret only one image. Asking one person to interpret one image and another person to interpret the other image complicates the comparison of interpretations. As two people might differ in their interpretation of one image, a different interpretation of the two images might be a result of individual interpretation differences or a result of the differences between the two images. Consider the descriptions in Table 3.1, which were generated by people who were asked to provide their interpretation of either Figure 3.2(a) or (b), to get an idea of the challenge that a comparison of subjective interpretations poses.

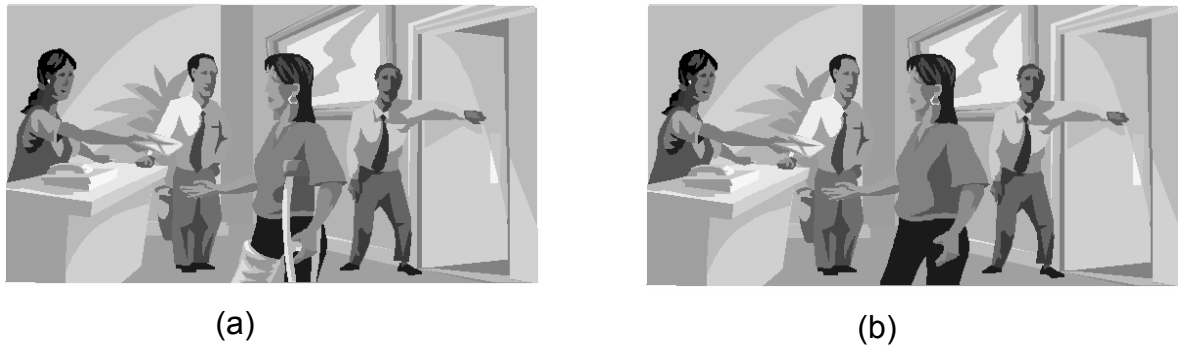


Figure 3.2. Example of a change possibly affecting the gist of the image. The deleted cast and crutch possibly result in different interpretations of images (a) and (b).

Table 3.1. Descriptions generated by participants who were asked for their interpretation of either Figure 3.2(a) or (b).

Descriptions of Figure 3.2(a)	Descriptions of Figure 3.2(b)
<ul style="list-style-type: none"> • A woman gets a letter at the reception of an office • In a hospital the receptionist gives a letter to a patient • Waiting room of the general practitioner • Check-up at the doctor's • Colleagues are happy their colleague is back at work, only she is not happy, probably because of her injury • Hospital • The desk at a family doctor's practice • People at the office kindly converse with the disabled woman • Joann with her broken leg gets redundancy pay • A patient gets a letter in the hospital • A woman is being assisted • A woman with crutches gets an envelope and has to leave the room • A woman with a broken leg receives an envelope from a woman behind the counter (Maybe in the hospital) • A woman with an injury on her left leg is assisted at the counter of a shop, while a man is opening the door for her 	<ul style="list-style-type: none"> • Someone receives a letter and then has to leave • A desk clerk who hands over an envelope to a woman • Activity at the reception • At the office a man opens the door for a woman • A woman gets a letter from a woman behind a desk and there are two men, one of whom opens the door for her • A woman gets an invitation • Getting fired • An office where something is handed over to a woman in front of a counter • People who are working at an office • Two men see a beautiful woman and look at her • Woman delivering an envelope to another woman • It is busy at the office

To determine whether Figure 3.2(a) and (b) are systematically interpreted differently, the descriptions of Figure 3.2(a) have to be compared with the descriptions of Figure 3.2(b). One possible method for comparing the descriptions in Table 3.1 would be to categorize them according to their semantic content. Generally, a categorization based on the content of the descriptions would require an extensive coding scheme with one category for every possible interpretation of the image and (potentially complex) rules for mapping descriptions onto

categories. More importantly, the coding scheme is prone to researcher bias. The categories, which the researcher decides to include in the coding scheme, could bias the interpretation of the descriptions by the person coding the descriptions in the direction of the researcher's hypotheses. For instance, the choice of included categories dictates whether descriptions mentioning "a patient" should be categorized differently than descriptions mentioning "a woman". Most likely the coding scheme will be developed, or at least adjusted after the collection of descriptions, since it is very difficult to anticipate on every possible interpretation. Another disadvantage of a categorization based on the semantic content of the descriptions is that each set of images requires a new coding scheme.

To overcome these shortcomings an alternative method for categorizing and comparing descriptions is proposed. In the proposed method only one criterion is used as the base for the categorization: an interpretation either fits or does not fit an image. The method makes it easier to determine whether two images are systematically interpreted differently.

3.2 Proposed method to determine a gist change

The method proposed in this section was designed to determine whether people would interpret two images differently. A systematic difference between interpretations of images can be determined in three steps. In the first step for each image a set of descriptions is generated by separate groups of people (the Generators). As the gist of an image is subjective, each image is interpreted by a number of Generators. By having several people interpret each image, a wide range of possible interpretations will be present in the set of descriptions. In the second step a person (the Rater) judges a "good fit" or a "poor fit" for each description to the images, providing the base for a categorization. In the third step the categorized descriptions are analyzed statistically. By analyzing the descriptions based on the assigned binary responses by the Raters, a decision of whether or not two images are described differently is less dependent on the generated descriptions themselves or the interpretation by the Rater. The method makes it possible to determine whether two images are interpreted differently, without the need to know exactly how each image is interpreted. The method is illustrated by applying it to the three circles of different sizes shown in Figure 3.3.

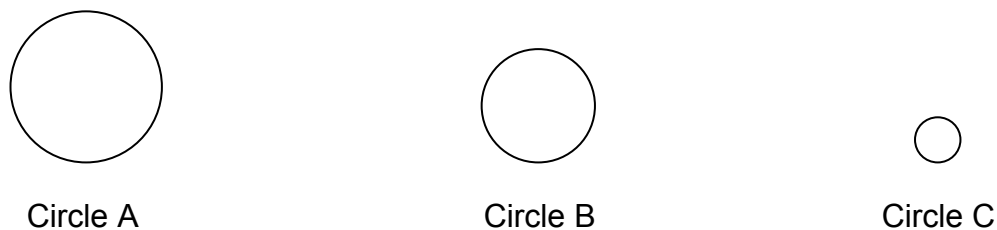


Figure 3.3. Three circles differing in size.

Step 1: collecting descriptions, the Generator task

The first step in determining if images are interpreted differently is to have human participants generate descriptions reflecting their interpretations. Because seeing more than one of the images could influence the way an image is described (Garner, 1974), for each image a separate group of people (the Generators) is asked to provide the descriptions. For the circle

example ten hypothetical descriptions of each of the circles from Figure 3.3 are presented in Table 3.2 below. As circles A and B are both described as being Large and Small about half of the time, it seems fair to say that the sets of descriptions of circles A and B are similar. Circle C on the other hand is only described as being Small. The descriptions of circle C can be said to be different from the descriptions of circles A and B. A difference or similarity in sets of descriptions is not always as obvious as in the circle example. Consider for instance the descriptions in Table 3.1 for the images in Figure 3.2. The descriptions in Tables 3.1 and 3.2 represent subjective interpretations of the images. Therefore, there is no qualitative difference between the descriptions, i.e., there are no "right" or "wrong" descriptions.

Table 3.2. Frequency of hypothetical descriptions Small and Large for the three differently sized circles in Figure 3.3.

Description	Circle A	Circle B	Circle C
Small	4	5	10
Large	6	5	-

Step 2: categorizing descriptions, the Rater task

In the second step participants (the Raters) who were not involved in generating the descriptions, and were naive to the nature of the task, judge the descriptions. Each description is judged based on whether the description could be an appropriate interpretation (good fit, 1) or a less appropriate interpretation (poor fit, 0) for each of the images. The Rater reads all descriptions and views all images. The Rater has no knowledge which description was generated for which image. The Rater's task results in a fit-response for each description for each image.

Considering the circles in Figure 3.3 and their descriptions there are four logical³ views a Rater could have concerning the appropriateness of the descriptions. A Rater could view all three circles as being Large (Hypothetical Rater 1), or all three circles as being Small (Hypothetical Rater 2). Alternatively, a Rater could consider circles A and B both to be Large and circle C to be Small (Hypothetical Rater 3), or circle A to be Large and both circles B and C as being Small (Hypothetical Rater 4). The interpretation of the Rater is reflected in the assigned fit-responses for each description for each image. For instance, Hypothetical Rater 1 will judge all descriptions Small to be a poor fit (0) for all three circles, and all descriptions Large to be a good fit (1) for each circle. In Table 3.3 below the fit-responses for the descriptions Small and Large matching the interpretations of the four Hypothetical Raters are displayed.

³ Other views are possible, such as: perceiving circle A and B "Small" and circle C "Large". These other options and their corresponding fit-responses will lead to the same conclusions as the discussed options.

The same way that it could be concluded that there is no qualitative difference between the descriptions, it could be concluded that there is no qualitative difference in interpretations by the Rater, i.e., there is no "right" or "wrong" interpretation of the descriptions by a Rater. All four interpretations as displayed in Table 3.3 are equally valid, and they should lead to the same conclusions about a systematic difference between sets of descriptions. Therefore, a conclusion on a systematic difference between sets of descriptions cannot be based on the fit-responses alone. For example, if a conclusion would be based on whether a description would fit equally to circle A and B, the interpretation of Hypothetical Rater 4 would lead to a different conclusion than the interpretations of Hypothetical Raters 1, 2 and 3. The fit-responses contain information about how a Rater interprets the images, while the objective of the method was to compare sets of descriptions.

Table 3.3. Fit-responses for descriptions Small and Large for the three circles from Figure 3.3, matching the interpretations of four Hypothetical Raters.

Description	Rater 1 (A+B+C Large)			Rater 2 (A+B+C Small)			Rater 3 (A+B Large, C Small)			Rater 4 (A Large, B+C Small)		
	A	B	C	A	B	C	A	B	C	A	B	C
Small	0	0	0	1	1	1	0	0	1	0	1	1
Large	1	1	1	0	0	0	1	1	0	1	0	0

Step 3: comparing sets of descriptions, the researcher's task

To determine whether the sets of descriptions of images are systematically different, the fit-responses are combined to represent a fit-pattern. When comparing the descriptions of two circles, each description can be denoted as a fit-pattern, combining the fit-responses for the circles. For an arbitrary pair of objects X and Y there are four possible fit-patterns for each description:

- (00) the description is not appropriate for both object X and object Y,
- (10) the description is appropriate for object X but not appropriate for object Y,
- (01) the description is not appropriate for object X but appropriate for object Y and
- (11) the description is appropriate for both object X and object Y.

A decision on a difference between sets of descriptions is not based on the observed fit-patterns. A frequent occurrence of the fit-patterns "10" and "01" is not necessarily proof of a difference between sets of descriptions of the images. The fit-patterns "10" and "01" indicate that the Rater interprets the two images differently. Hypothetical Rater 4 in Table 3.3 illustrates that the occurrence of fit-patterns "10" and "01" does not necessarily mean a difference between sets of descriptions. Although Hypothetical Rater 4 discriminates between circles A and B, this does not affect the similarity between the sets of descriptions of circle A and B. A fit-pattern reflects the appropriateness of a description for circles A and B based on the Raters interpretation of the description and the circles. A difference or similarity between

the sets of descriptions of circles A and B depends on the number of times each circle is described as being Large or Small.

The source of a description (circle A, B or C) should be considered to find a systematic difference in occurring fit-patterns between the sets of descriptions. Table 3.4 contains the hypothetically occurring fit-patterns for the descriptions of circles A, B and C. The columns of this table represent the source of the descriptions (circle A and B in Table 3.4(a) and circle A and C in Table 3.4(b)). The rows represent the fit-patterns (00, 10, 01 and 11). In the cells the number of descriptions which were appointed the fit-pattern is displayed. Consider for instance the fit-patterns for descriptions of circles A and B for Hypothetical Rater 1 (first column in Table 3.4(a)). As Rater 1 interprets both circle A and circle B to be Large, all descriptions "Small" are appointed the fit-pattern "00", and all descriptions "Large" are appointed the fit pattern "11". Circle A was described as being Small 4 times. This results in a "4" in the cell indicating the "00" fit-pattern for the descriptions generated for circle A. Circle A was also described as being Large six times. This results in a "6" in the cell indicating the "11" fit-pattern for the descriptions generated for circle A. Circle B was five times described as being Small and five times described as being Large. For Rater 1 this resulted in a "5" in both the "00" and "11" cells for the descriptions generated for circle B.

Table 3.4(a). Frequency of hypothetically occurring Fit-patterns for descriptions of circle A and B for four Hypothetical Raters.

Fit-pattern	Source of Descriptions (Circle A and Circle B)*							
	Rater 1 (A+B+C Large)		Rater 2 (A+B+C Small)		Rater 3 (A+B Large, C Small)		Rater 4 (A Large, B+C Small)	
	A	B	A	B	A	B	A	B
00	4	5	6	5	4	5	-	-
10	-	-	-	-	-	-	6	5
01	-	-	-	-	-	-	4	5
11	6	5	4	5	6	5	-	-

* Descriptions of circle A: Small (4x), Large (6x)
 Descriptions of circle B: Small (5x), Large (5x)

Table 3.4(b). Frequency of hypothetically occurring Fit-patterns for descriptions of circle A and C for four Hypothetical Raters.

Fit-pattern	Source of Descriptions (Circle A and Circle C)*							
	Rater 1 (A+B+C Large)		Rater 2 (A+B+C Small)		Rater 3 (A+B Large, C Small)		Rater 4 (A Large, B+C Small)	
	A	C	A	C	A	C	A	C
00	4	10	6	-	-	-	-	-
10	-	-	-	-	6	-	6	-
01	-	-	-	-	4	10	4	10
11	6	-	4	10	-	-	-	-

* Descriptions of circle A: Small (4x), Large (6x)
 Descriptions of circle C: Small (10x)

If there is a systematic difference in the way images are described, then there should be a difference between the fit-patterns for the associated sets of descriptions. When the fit-patterns assigned to the descriptions of one image differ from those assigned to the descriptions of another image, this indicates a difference between the sets of descriptions. A difference does not depend on *which* fit-patterns occur. In the circle example, the different interpretations of the Hypothetical Raters resulted in different fit-patterns. Note that this should not affect the conclusion, as all Raters have seen the same descriptions, and all Rater interpretations are equally valid. So, it is not the fit-patterns, but the *distribution* over the fit-patterns that determine whether there is a systematic difference between sets of descriptions. Statistical analysis will reveal whether two fit-pattern distributions are significantly different.

Statistical analysis

As already suggested by the hypothetical data in Table 3.4, in reality several fit-patterns occur rarely, or not at all. Conventional tests for contingency tables (such as a Chi-square test) will be unreliable or produce no results at all with sparsely-filled tables. For this case, a Fisher-Freeman-Halton Exact probability test for contingency tables can be used (Freeman and Halton, 1951). The Exact test calculates the sum of the probability of occurrence the frequency distribution of the specific table and the probabilities of all frequency distributions that are less probable (i.e., "more extreme"). If the sum of these probabilities is less than 0.05 it is concluded that there is a significant difference between the fit-pattern distributions for descriptions of two images. We may then conclude that this indicates a systematic difference between sets of descriptions. An analysis of fit-pattern distributions for the circle example in Table 3.4(a) reveals no difference between the fit-pattern distributions for the descriptions of circles A and B ($p = 1.0$ for all four Hypothetical Raters). Analysis of the data in Table 3.4(b) reveals a difference between the fit-pattern distributions in the descriptions of circles A and C ($p = 0.011$, for all four Hypothetical Raters). Based on these results we conclude that circles A and B are described in a similar way and circle C is described in a different way than circle A.

3.3 Validity of the method

In this section the validity of the proposed method and some of its benefits and potential drawbacks are considered. Finding a significant difference between fit-pattern distributions when sets of descriptions of two images are similar is unlikely. As demonstrated in the circle example, even if a Rater interprets circle A and B as being different (Hypothetical Rater 4), this does not lead to a different conclusion for a difference between the sets of descriptions. The only way a Rater can produce fit-patterns indicating a difference in descriptions if there is in fact no difference, is when a Rater is extremely lucky in guessing the source of a description and judges accordingly. In the circles example this would mean that a Rater would discriminate between descriptions denoted as "Large" generated for circle A and descriptions denoted as "Large" generated for circle B. Not only is it unlikely that a Rater would be able to make this distinction, it also means that the Rater would be inconsistent as he or she needs to assign different fit-responses for identical descriptions. The chance of finding a significant difference between two fit-pattern distributions when there is in fact no difference in descriptions (type I error) is small.

The chance of finding no difference between fit-pattern distributions when sets of descriptions of two images are different is relatively large (type II error). The method is conservative.

Raters have to judge if a description would be appropriate for an image. A Rater with a lot of imagination might feel that all descriptions are appropriate for each image (e.g., descriptions Large and Small are equally appropriate for all three circles). In this case, such a liberal interpretation would result in all "11" fit-patterns, resulting in no difference between fit-pattern distributions. Similarly, a Rater lacking imagination might conclude none of the descriptions is appropriate, resulting in all "00" fit-patterns. In such extreme cases, even when one circle is consistently described as Small and the other described as Large, no difference between fit-pattern distributions will be found. Consequently, when the fit-pattern distributions are not significantly different, this does not imply that the descriptions are the same. The method will detect a gist change if it is clear to both Generators and Raters that there is a difference.

The method becomes more reliable with more descriptions for each image. With more descriptions per image the interpretation of a larger group of people is represented, making it more likely that the set of descriptions reflects the general interpretation of an image for a particular target group. Also, with more descriptions, a less common interpretation of an image will have correspondingly less weight in the final analysis. The minimum number of descriptions generated for each image to assess a difference in interpretation between two images may vary with the complexity of the images. Complex or ambiguous images may require more descriptions than images displaying a univocal scene. Although the necessary number of descriptions was not assessed directly, the recommended number ranges between 10 and 20 descriptions for each image.

Asking several Raters to judge the descriptions can increase confidence for the researcher. When using more than one Rater, the data from each Rater is analyzed separately. As demonstrated in the circle example the fit-patterns provided by the Raters do not need to be the same to reach the same conclusion. Differences can occur when Raters have varying interpretations. When different Raters reach different conclusions this can be the result of three reasons. First, a Rater might have been extremely lucky in guessing and has produced fit-patterns indicating a difference even though the sets of descriptions do not differ (type I error). Second, a Rater might not share the interpretations of the Generators (type II error). Third, the sets of descriptions may not be sufficiently different. When the difference between sets of descriptions is small, different fit-patterns for only a few descriptions can result in different conclusions for different Raters. As the chance that a Rater produces data indicating a difference in descriptions when there is in fact no difference in descriptions is small, the chance that two independent Raters produce such data for the same sets of descriptions is negligible. Therefore, when two Raters produce significant differences, there is evidence that sets of descriptions are systematically different. As the chance of finding no difference between fit-pattern distributions when in fact the sets of descriptions are different is relatively large, using three Raters is recommended. When the data from two out of three Raters produces a significant outcome, this indicates a gist change. For an even more conservative approach, agreement between all three Raters can be applied as criterion.

Conclusion and Implications

In this chapter a method is proposed to determine whether a change in an image affects the gist of that image. The gist of an image is reflected in a verbal description of a subjective interpretation on what the image is about. To determine a gist change, subjective

interpretations of the original image need to be compared to interpretations of the altered version of the image. The method overcomes some of the problems associated with categorizing and comparing sets of subjective interpretations. Only one criterion is used as the base for the categorization: an interpretation either fits or does not fit the image. It is concluded that sets of descriptions are systematically different if there is a difference between fit-pattern distributions for two sets of descriptions.

The proposed method makes it possible to compare subjective interpretations of images, while leaving room for differences in interpretation of an image by different people. Compared to categorizing descriptions based on their semantic contents, the proposed method of categorizing based on fit-patterns has two advantages. First, the categorization is not prone to researcher bias, as the researcher does not need to define categories. Second, the fixed categories make the method applicable to all kinds of images, without the difficulty of coming up with extensive coding schemes.

A legitimate concern is that a gist change, as determined with the proposed method, need not necessarily reflect a gist change for everyone. As knowledge, experience, interests and expectations influence the interpretation of images, descriptions will reflect the subjective interpretations for a particular group of Generators⁴. Strictly speaking, the proposed method makes it possible to determine if a particular target group, in this case Dutch university students, would perceive a change as affecting the gist of an image. However, this does not greatly limit the applicability of the method. If this method is used to select images for further research, such as research on the effect of gist change on image recognition or difference detection, the target group in the ensuing studies should be chosen identical to the target group used for the image selection.

The method described is not limited to the detection of gist changes in images. By comparing the results from the method for two different target groups, for instance experts and novices, a possible difference in image interpretation can be found. At a more general level, the method might be applicable to determine any significant difference in subjectively interpretable stimuli (taste, smell or even associations with brand names). The (only) two conditions that need to be fulfilled are that a description of the interpretation of the stimuli can be generated and that these descriptions can be judged by one or more Raters as "likely" or "not likely" descriptions of that stimulus.

⁴ Note that the Raters do not necessarily have to be part of the same target group as the Generators. The Raters can only produce fit-pattern distributions which indicate a difference between sets of descriptions if there is indeed a difference between the sets. When a Rater does not share the interpretation of the target group this can only lead to data indicating no difference (i.e., type II error). To prevent type II errors preferable Raters should be used that share the interpretations of the particular target group.

4

Validation of test stimuli⁵

In this thesis the role of gist in memory for images is addressed. In chapter 2 the gist of an image was defined as a subjective interpretation of what an image is about. Chapter 3 proposed a method to determine whether two images have a different gist or not. In chapters 5 and 6 experiments are reported that aim at testing the hypothesis that people are more sensitive to changes affecting the gist compared to changes that do not affect the gist. These experiments require carefully developed stimuli that fulfill the criteria needed to test the hypothesis. This chapter addresses the criteria, presents the developed stimuli and explores the validity of the stimuli.

4.1 Considerations for valid test images

In the next two chapters experiments are reported that set out to test sensitivity to changes affecting the gist compared to sensitivity to other changes. The sensitivity is measured using a memory difference-detection task. The general setup of the experiments is as follows. There is a study phase in which several images are studied, each for 5 seconds. In a test phase the memory for the studied images is tested by asking participants whether or not the image in the test phase is exactly the same as the image in the study phase. To determine a difference in sensitivity for gist and other image properties, performance for two types of images will be examined.

- Gist Change images: images similar to the studied image except for one item, which affects the gist of the image.
- Feature Change images: images similar to the studied image except for one item, which does not affect the gist.

A greater sensitivity for gist compared to sensitivity for other image properties can be determined by comparing the detection of the changed items for the two types of images. To

⁵ Preliminary results of the research described in this chapter were presented at the 5th annual meeting of the Vision Sciences Society (VSS, 2005).

be able to compare change detection performance for Gist Changes and other changes (referred to as Feature Changes), the applied changes should be perceived as either affecting the gist or not affecting the gist, by the participants in the experiment. As gist is a subjective interpretation of an image, a change can affect the gist for one group of people, but not for another. For instance, for disabled people the wheelchair in Figure 4.1 might be part of the gist, whereas for younger people the image might be interpreted as an old person and a computer.

For a direct comparison between the two types of changes, preferable the only difference between the two types of changes should be whether or not the gist is affected. For instance, removing the computer screen in the image in Figure 4.1 would surely affect the interpretation of the image (i.e., the gist), and removing the collar of the person sitting in front of the computer will probably not affect the gist of the image. However, a difference in detecting the missing computer screen compared to detecting the missing collar would not necessarily reflect a difference in sensitivity to gist compared to other image properties. To name only one possible alternative explanation, a difference in detection of the two changes could be based on the size of the affected area.



Figure 4.1. Although deletion of the computer screen would be a change affecting the gist, the change would be confounded with other factors affecting perceptibility, such as size.

The validity of the test stimuli depends on three criteria.

- The changes should not be confounded with other factors, which might facilitate the detection of Gist Changes more than the detection of Feature Changes.
- The Gist Change should be perceived as a change affecting the gist by the target group.
- The Feature Change should not affect the gist as it is perceived by the target group.

The last two criteria are addressed in the next section. In this section four possible confounds are discussed briefly: the type of image, the size of the change, the position of the change in the image, and the type of change. This section ends with the presentation of the developed test stimuli.

Four possible confounding factors

The first possible factor that could affect the perceptibility of a change is the type of image. Some images might be easier to remember than other images as a result of personal interest or the complexity of an image. Changes in images that are easy to remember might have an advantage in a change detection task. To ensure that a possible difference in change detection between Gist Changes and Feature Changes is not confounded with memory for the whole image, both types of changes are applied to the same image. Thus, for each original image a Gist Change image and a Feature Change image were developed.

The second possible factor which could affect the perceptibility of a change is the size of the area affected by the change. All else being equal, changes affecting a large area of an image will be more visible than changes affecting a small area. When a change affects a large area, there is a large chance that the area will be fixated upon, enabling detection of the change. Also, large (pictorial) changes are more likely to disrupt the layout of an image. To ensure that a possible detection advantage for Gist Changes is not confounded with the size of the affected area, the Gist Change on average comprised a smaller portion of the original image than the corresponding Feature Change (see Table A.1 in the appendix).

The third possible factor which could affect the perceptibility of a change is the position in the image where the change occurs. From research on paintings perception (Locher, Gray, & Nodine, 1996) it is known that people pay more attention to the centre than to the edges. In images, just as in paintings the centre is usually the most informative area. This has implications for the possibilities of Gist and Feature changes. As much of the information of an image is available in the centre, a change affecting the gist is more likely to be positioned in this informative area. Consequently, a Feature change is rarely positioned in the centre, because a Gist and Feature Changes never occurred at the same position. The position in the image where the change occurs was varied as much as possible between the test sets. Figure A.2 in the appendix contains an overview of the position of the changed area.

The fourth possible factor which could affect the perceptibility of a change is the type of change. Two types of changes are addressed: physical changes and semantic changes. Physical changes relate to properties of the image (e.g., color, size, position, and spatial layout). Differences in visibility of physical changes are assumed to be caused by differences in the speed of perception and encoding for different physical properties (see for instance Aginsky and Tarr, 2000 and Tatler, Gilchrist and Rusted, 2003). Semantic changes relate to the meaning of an image. The changed object or property has a particular relation to the meaning of the image. For instance, in an image of a kitchen a coffee maker is a consistent object. In an image of a bathroom, the same coffee maker is an inconsistent object. Differences in visibility between different types of semantic changes are assumed to be caused by the amount of attention the object or aspect receives during the study phase (Friedman, 1979). Examples of studies that have reported on semantic changes, are change detection for objects or areas of central interest (CI) and marginal interest (MI) (Rensink et al., 1997), semantic and non-semantic changes (Werner and Thies, 2000), and consistent and inconsistent objects (Hollingworth and Henderson, 2000; Biederman, 1981). Although semantic changes are closely related to Gist Changes, a semantic change is not necessarily the same as a Gist Change. For instance, in an image displaying a doctor with a patient, substituting the doctor (which is an object of central interest) by another doctor does not necessarily affect the interpretation of the image (test set 12 in Figure 4.2). The image with the new doctor can also

be interpreted as an image about a doctor with a patient (see also Simons & Levin, 1998). To prevent confounding of types of changes, the types of changes applied in the test images were varied as much as possible. An overview of the changes applied to each image is presented in Table A.1 in the Appendix of this thesis.

Developed test stimuli

To serve as test stimuli colored Microsoft Clipart images were selected. The selected Clipart images are mostly more complex than line drawings, but less complex than photos. The use of drawings allowed for addition of mythical features such as Dracula teeth and a Loch Ness monster without distorting the consistency of the scene (scenes 1 and 13, Couple and Campfire in Figure 4.2). As stated above, to prevent confounding of the type of image both the Gist Change and the Feature Change are applied to the same image. This resulted in sets of images to be tested. Each test set consists of an Original image, a Gist Change image, and a Feature Change image.

Besides the Original, Gist Change and Feature Change image, for each Original image another clipart image was selected. In chapter 1 gist was defined as: a personal interpretation of what an image is about. As said in chapter 2, based on Action Identification Theory (Vallacher & Wegner, 1987) it is expected that people will interpret an image at the highest possible level that still contains enough detail to be meaningful. This preferred level of interpretation might be the level of semantic category. If this is the case, the gist corresponds to the commonly used definition of gist: a short description capturing the identity of a scene (e.g., Hollingworth & Henderson, 2002). By adding a Thematically Related image to each test set, it is possible to gain some insight in the similarity between the commonly used definition of gist, and the definition used in this thesis.

Eighteen test sets were developed, each containing four types of images, an Original image, a Gist Change image, a Feature Change image, and a Thematically Related image. The types of images are addressed separately below. All 18 test sets are displayed in Figure 4.2.

The Original image

The Original images were 18 colored Microsoft clipart images (first column in Figure 4.2). The images were selected from an extensive collection of Microsoft Clipart images. The images depicted a variety of in- and outdoor scenes and varied in complexity. The images contain persons (e.g., just two people), situations (e.g., a car crash, with a wounded person being attended to in the foreground), and sceneries such as mountains and a lake.

The Gist Change image

A Gist Change is a change that influences the possible interpretation of an image. For each of the 18 Original images, there is a Gist Change image (second column in Figure 4.2). The changes are meant to alter the interpretation of what the image is about. The author interpreted the Original images by asking questions like: what does the image want to make clear, who is doing what, where is it situated. Based on these interpretations alterations were made to change the answer to one of the questions used for the interpretation of the image. In test set 1, (Couple) in Figure 4.2 for instance, an image interpreted as "two people showing affection" changes to an image which could be interpreted as "a man intending to harm a woman". By

altering location cues, such as the Eiffel tower or kangaroos, the answer to the question "where is it situated" can be changed (test sets 10 and 14, Paris and Vacation in Figure 4.2).

The Gist Change images were created by altering one aspect or object in the original image. With Paint Shop Pro in each of the 18 original images an object was deleted, added, mirrored or substituted by a different object. The added and substituted objects were placed in the image in such a way that their sizes and positions fit seamlessly in the scene.

The Feature Change image

A Feature Change is a change that does not influence the possible interpretation of an image. For each of the 18 Original images, there is a Feature Change image (third column in Figure 4.2). Feature changes modify the color or add an object, and leave the gist of the image intact. For instance, in test set 1, (Couple) in Figure 4.2, an image interpreted as "two people showing affection" is probably not interpreted differently by a switch in hair color.

To accomplish a Feature Change, the author interpreted the Original image and an alteration was made that did not change the answer to a question used for the interpretation. The Feature Change images were created by altering one object or property in the original image. Using Paint Shop Pro objects were added, deleted, mirrored, substituted or changed in color. The altered objects were placed in the image in such a way that their sizes and positions fit seamlessly in the scene.

Thematically Related image

The Thematically Related images were new Clipart images, selected to match the central theme (e.g., Couple) of the Original image, but to be pictorially clearly different (last column in Figure 4.2).

In this section three criteria for valid stimuli were stated. To meet the condition that the changes should not be confounded with other factors that might facilitate the detection of Gist Changes more than the detection of Feature Changes, four possible confounding factors were addressed. The developed test stimuli were presented. The next section addresses the remaining two criteria.

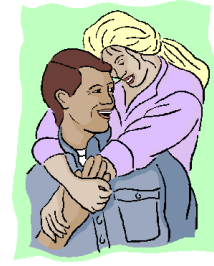
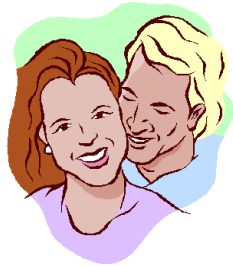
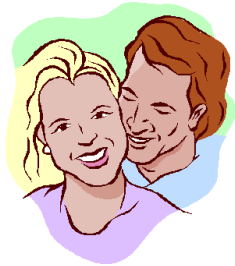
Original

Gist Change

Feature Change

Thematically Related

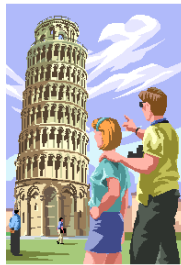
1. Couple



2. Bar



3. Pisa



4. Realtor



5. Jogging



6. Hospital



Figure 4.2 continues on next page

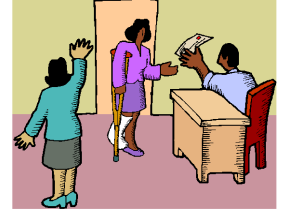
Original

Gist Change

Feature Change

Thematically Related

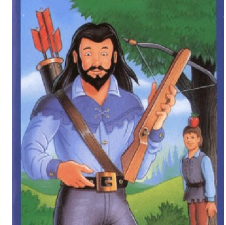
7. Reception



8. Seine



9. William Tell



10. Vacation



11. New year



12. Doctor



Figure 4.2 continues on next page

Original

Gist Change

Feature Change

Thematically Related

13. Campfire



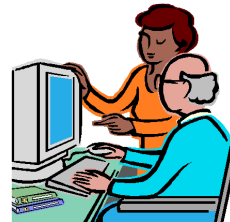
14. Paris



15. Gambling



16. Computer



17. Car crash



18. Still life



Figure 4.2. 18 test sets consisting of the Original image, a Gist Change image, a Feature Change image and a Thematically Related image. The validity of the labels is determined in section 4.2.

4.2 Determining validity of the changes

In this section the remaining two criteria for valid test stimuli are addressed.

- The Gist Change should be perceived as a change affecting the gist by the target group.
- The Feature Change should not affect the gist as it is perceived by the target group.

To determine if the gist of an image is affected as a result of a change, the interpretations of the Original image and the changed image have to be compared. If the Original image is systematically interpreted (described) differently than the changed image, the gist of the image is affected as a result of the change. For validity of the Feature Change, the Feature Change image should be interpreted similar to the Original image. Whether the Gist Change images are interpreted systematically different from the Original image and the Feature Change images are interpreted similar to the Original image is empirically determined by applying the method described in chapter 3. The method consists of three steps. In the first step descriptions are collected for each image. In the second step the descriptions are judged on whether or not they could fit each image. In the third step the judgments for the descriptions are analyzed statistically.

Step 1: collecting descriptions, the Generator task

The image descriptions were generated by 154 students of the University of Utrecht (106 female and 48 male, average age 20.9 year) and 59 students of the Radboud University Nijmegen (34 female and 25 male, average age 20.0 year). The Generator task was part of a set of unrelated tasks, presented on a laptop computer. All tasks were programmed in Authorware 5. Participants were paid 5 Euros upon completion of all the tasks.

The Generators were informed that the purpose of the research was to gain insight into how people interpret images. Each Generator was presented with one image from a particular test set. So a Generator saw either the Original, Gist Change, Feature Change or Thematically Related image of a particular test set. The image was shown on the computer screen for five seconds. The Generator was asked to type in a text field his or her interpretation of what the image was about. When a participant had finished typing the description he or she clicked a "next" button to continue with the next image. Each Generator described five images.

The participants from the University of Utrecht described 15 test sets. As each image from a test set was described by a different person, collecting descriptions for each image in a test set requires four participants. Each participant generated descriptions for five images. Consequently, 12 participants were required to provide one description for each of the test stimuli in the 15 test sets. Participants were randomly assigned to one of 12 conditions. The participants from the Radboud University Nijmegen described the remaining three test sets in four conditions. The assignment of participants to experimental conditions was unbalanced, resulting in different numbers of descriptions generated for each image. The number of descriptions generated for each image varied between 10 and 16.

Step 2: categorizing descriptions, the Rater task

The descriptions were rated by three staff members (1 female, average age 31.3 years), naïve to the nature of the task. Each Rater received a booklet containing 18 rating tasks (one task per test set). Each new task started on a new page. An example of part of a Rater task for one test set is displayed in Figure 4.3.

The rating task consisted of all four images of a given test set side by side, followed by all descriptions generated for these images in random order. The Raters were told that the people who provided the descriptions had been asked to describe their interpretation of the image. The Raters were told to: (i) look at the images, (ii) notice the differences, and (iii) judge for each description whether it was a good fit (1), or a poor fit (0) for each of the images.

It was stressed that the Rater should consider whether someone could possibly have given the description and not whether the description reflected the Rater's personal interpretation. Although the method was designed to be able to handle different interpretations by different people (see chapter 3), it is possible that a Rater's personal interpretation is not reflected in the generated descriptions. If a Rater would rate the descriptions only based on his or her own personal interpretation, then the chance of finding no difference between sets of descriptions increases. If for instance a Rater's personal interpretation of the images in Figure 4.3 would include a mentioning of the fact that the woman is wearing a pearl in her ear, allowing the Rater to judge the appropriateness of the descriptions only based on his or her own interpretation would result in "poor fit" judgments for all descriptions which do not mention the earring. Based on the results of such a critical Rater, it would then be concluded that there is no difference between descriptions.

Finally the Rater was asked to be as consistent as possible. The Rater's task resulted in four fit-responses for each description, one for each image.





				
Love	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit
Loving	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit
Portrait of a couple	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit
Scary love affair	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit
Symbolic representation of a happy western couple	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit
People who like each other	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit
Vampire bites woman	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit
A man and a woman, probably a couple. he tells her something	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit
Undercover Dracula	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit
A man and woman, probably in love, in some kind of photo	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit
A vampire bites the neck of an ignorant woman.	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit
Showing affection	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit	<input type="checkbox"/> Good Fit <input type="checkbox"/> Poor Fit

Figure 4.3. Example of part of a Rater task for one test set. The images and descriptions each generated for one of the four images were displayed in random order. The Rater judges for each description whether it fits each image.

Step 3: comparing descriptions, the researcher's task

For the analysis the descriptions with their assigned fit-responses were sorted according to the source of the description. The descriptions are grouped based on the source of the descriptions, resulting in four groups of descriptions for each Rater for each of the 18 test sets. The fit-responses were analyzed for each Rater separately. For determining whether the Original images were described in a systematically different way from the Gist Change images, only the descriptions of those two images were considered. For the two groups of descriptions the fit-responses for the two images were combined into a fit-pattern. An example of the fit-responses for some of the descriptions for a complete test set is displayed in Table 4.1. For the comparison of the Original image and the Gist Change image the fit-responses for the other two images from the test set (grey part of Table 4.1) were disregarded.

Table 4.1. Example of fit-responses for each of the four images for four groups of descriptions for one Rater. For the comparison of the descriptions of the Original image and the Gist Change image the data in the grey area are disregarded.

Descriptions of Original	Original	Gist Change	Feature Change	Thematically Related
- a man and woman, probably a couple. he tells her something	1	0	1	0
- loving	1	0	1	1
- portrait of a couple	1	0	1	1
- a man and woman, probably in love, in some kind of photo	1	0	1	1
Descriptions of Gist Change				
- scary love affair	0	1	0	0
- vampire bites woman	0	1	0	0
- a vampire bites the neck of an ignorant woman.	0	1	0	0
- Undercover Dracula	0	1	0	0
Descriptions of Feature Change				
- symbolic representation of a happy western couple	1	0	1	1
- Love	1	0	1	1
Descriptions of Thematically Related				
- showing affection	1	0	1	1
- People who like each other	1	0	1	1

When comparing the descriptions of the Original and Gist Change image there are four possible fit-patterns:

- (00) the description does not fit the Original image and the Gist Change image,
- (10) the description does fit the Original image but not the Gist Change image,
- (01) the description does not fit the Original image but fits the Gist Change image,
- (11) the description does fit the Original image and the Gist Change image.

The occurring fit-patterns for the two groups of descriptions can be presented in a contingency table. The columns of this table represent the source of the descriptions (Original image and Gist Change image), the rows represent the fit-patterns (00, 10, 01 and 11). The cells contain the number of descriptions that were assigned a particular fit-pattern. The data from each Rater is treated as a separate contingency table. For an example of a similar contingency table see Table 3.4 in chapter 3.

To compare the descriptions of the Original image to descriptions of the Feature Change image and to descriptions of the Thematically Related image the same procedure applies. So, for a comparison between descriptions of the Feature Change image and descriptions of the Original image, the descriptions and fit-responses for the Gist Change image and the Thematically Related image are disregarded. Three comparisons are reported: Original image vs. Gist Change image, Original image vs. Feature Change image, and Original image vs. Thematically Related image. As the ratings are analyzed for each Rater separately, this resulted in nine contingency tables (three comparisons for each Rater). Each contingency table was analyzed using the Fisher-Freeman-Halton Exact probability test. The test indicates the probability that the distribution of the number of descriptions associated with each fit-pattern occur by chance (see chapter 3 for a more detailed discussion).

Results

Comparison of the fit-pattern distributions for descriptions of the Original and descriptions of the Gist Change image showed that for 11 sets the fit-pattern distributions are significantly different ($p < 0.05$) for each of the three Raters. For one test set the comparison showed a significant difference for two Raters, but not for the third Rater. For two test sets the comparison showed a significant difference for only one Rater. For the remaining four test sets the comparison showed no significant difference for any of the Raters.

Comparison of the fit-pattern distributions for descriptions of the Original and descriptions of the Feature Change image showed no significant difference for any of the Raters. For two test sets the comparison did show a marginal difference ($p < 0.10$) between the fit-pattern distributions for one of the Raters.

Comparison of the fit-pattern distributions for descriptions of the Original and descriptions of the Thematically Related image showed that for 11 sets the fit-pattern distributions are significantly different ($p < 0.05$) for each of the three Raters. For two test sets the comparison showed a difference for two Raters, but not for the third Rater. For two other test sets the comparison showed a difference for only one Rater. For the remaining three test sets the comparison showed no difference for any of the Raters. The results are summarized in Table 4.2 on the next page.

Table 4.2. Number of Raters (0,1,2,3) for which the comparison of the fit-pattern distributions for the descriptions of the Original image, the Gist Change, Feature Change and Thematically Related image showed a significant difference ($p < 0.05$).

Test set	Gist Change	Feature Change	Thematically Related
1. Couple	3	0	2
2. Bar	3	0	2
3. Pisa	3	0	1
4. Realtor	3	0	3
5. Jogging	3	0	3
6. Hospital	3	0	3
7. Reception	3	0	0
8. Seine	3	0	3
9. William Tell	3	0	0
10. Vacation	3	0	3
11. New year	3	0*	1
12. Doctor	2	0	3
13. Campfire	1	0	0
14. Paris	1	0	3
15. Gambling	0	0	3
16. Computer	0	0	3
17. Car crash	0	0	3
18. Still life	0	0*	3

* A marginal ($p < 0.10$) difference for one Rater.

Discussion

Valid test stimuli

A significant outcome for the comparison of the fit-pattern distributions for the descriptions of two images indicates that the descriptions of the two images are systematically different. This is interpreted as proof that the two images are interpreted differently, and therefore have a different gist. Because the test sets will be used in further experiments to examine the sensitivity for Gist Changes and Feature Changes the results are interpreted conservatively. Results from the comparisons between the interpretations of the Original image and the Thematically Related image held no consequences for the validity of a test set.

The label "Gist Change" was considered valid if the judgments from all three Raters resulted in a significant outcome. Therefore, it was concluded that for 11 test sets (test sets 1-11) the "Gist Change" label is valid.

The label "Feature Change" was considered valid when there is no indication for a difference between the descriptions of the Original image and the Feature Change image for any of the Raters. Although none of the comparisons for the Original image and Feature Change image showed a significant outcome, a marginal difference ($p < 0.10$) indicates there might be a

difference between sets of descriptions for the two images. Therefore, it was concluded that for all test sets except test set 11 and 18 the "Feature Change" label is valid.

A test set is considered valid if the images meet the two criteria; the Gist Change is valid and the Feature Change is valid. This was the case for 10 test sets (test sets 1-10 in Figure 4.2 and Table 4.2).

For the comparison between the descriptions of the Thematically Related image and the descriptions of the Original image a significant outcome can be interpreted to indicate that the descriptions of the two images contained sufficient information to distinguish the Original images from the Thematically Related images. If participants would interpret images at the level of the central theme of an image (i.e., a commonly used definition of gist), no difference between descriptions was expected. For only three Thematically Related images the descriptions are similar to the descriptions of the Original image. Looking at the descriptions provided by the Generators (see for instance Tables 3.1, 4.1 and 4.3) reveals that the interpretations of the Generators included more than just the category of an image (expectation stated in chapter 2). For example, for the test set Jogging (test set 5 in Figure 4.2) a lot of descriptions of the Original image referred to the relative position of the two people (descriptions motioned for instance "one behind the other" or "chase").

Invalid test stimuli

There are two possible situations in which two images that are described differently can lead to fit-pattern distributions that indicate that the two images are described equally.

- Situation 1. the two images are systematically described differently, but the Rater does not share the interpretations of the Generators,
- Situation 2. the two images are described differently by some Generators, and described similar by other Generators.

To clarify the two situations, the descriptions for the Original image and the Gist Change image for the test set Paris are provided below in Table 4.3. The appointed fit-patterns for each of the three Raters are provided in Table 4.4.

An example of the first situation (Rater does not share the interpretations from the Generators) can be found in the appointed fit-patterns for the descriptions of the Original image and the Gist Change image for test set Paris. Rater 1 judged almost all descriptions to fit both images, regardless of whether location (Paris) was mentioned in the description (see fit-patterns for Rater 1 in Table 4.4). Presumably, this Rater was familiar with Paris and recognized the Gist Change image as a typical Parisian scene, even without the Eiffel tower in the back. When a Rater does not share the interpretations of the Generators this could result in a type II error; concluding that the descriptions are the same, while they are not.

Table 4.3. Descriptions provided by Generators for the Original image and the Gist Change image for the test set Paris (test set 14 in Figure 4.2).

Descriptions of the Original image	Descriptions of the Gist Change Image
<ul style="list-style-type: none"> • Picturesque Paris • Vacation in Paris • Visiting Paris • A couple looks at the work of a Parisian painter • Parisian street scene • Man and woman on vacation in Paris, looking at the works from a landscape painter • A man is painting in Paris, two people are watching 	<ul style="list-style-type: none"> • Street drawings in Paris • Tourists in Paris watching an working artist • A street painter in France. • Painter working in a Mediterranean country. • Italy, maybe Firenze, painter, gets attention • A man painting and a married couple watching on a sunny day • A man and a woman watching a painter is painting in the street.
<ul style="list-style-type: none"> • Painters square in France • A French painter 	<ul style="list-style-type: none"> • Painting, vacation • A man painting, and two tourists watching the painting
<ul style="list-style-type: none"> • Painter makes painting while two people watch • Vacation • A man painting in the street and couple of tourists attentively watching the painting the painter is working on. • Man is painting 	<ul style="list-style-type: none"> • Painter on square with tourists • Paintings for sale • A man is painting outside, two people are watching. The man seems more interested than the woman who watches with her arms crossed.

Table 4.4. Frequency of occurring Fit-patterns for descriptions of the Original image and the Gist Change image for three Raters. The descriptions are presented in Table 4.3.

Fit-pattern	Source of Descriptions					
	Rater 1*		Rater 2**		Rater 3**	
	Original Image	Gist Change	Original Image	Gist Change	Original Image	Gist Change
00	-	1	2	-	-	-
10	-	-	9	3	7	3
01	-	1	-	3	-	2
11	13	10	2	6	6	7

* Presumed interpretation by the Rater: "All images are Paris".

** Presumed interpretation by the Rater: "Original image is Paris, Gist Change image is not Paris".

The second situation in which different descriptions of two images would not result in significant differences between fit-pattern distributions for the descriptions is when the descriptions are not sufficiently different. As the interpretation of images is subjective, it is possible that some people would interpret two images similarly, while others would interpret the images differently. This was also the case in the Paris test set. As can be seen from the descriptions in Table 4.3, some Generators describing the Original image did not mention any location cues. Also some Generators describing the Gist Change image did mention Paris.

Because of the limited difference in the descriptions of the two images (i.e., the sets of descriptions were not sufficiently different), a small difference in interpretation between two Raters (Raters 2 and 3 in Table 4.4) resulted in a significant outcome for one Rater, but not for the other.

The two situations described previously have direct implications for the validity of Feature Changes. As there are situations in which different descriptions of two images can lead to no significantly different fit-pattern distributions for the two images, a not significant outcome cannot directly be interpreted as proof that the two sets of descriptions are similar. For the research in this thesis two additional criteria are used to guard against concluding that images are described similarly while the similarity is only caused by the interpretation of the Rater. First of all for a decision on the validity of a Feature Change all three Raters have to agree. The possibility that all three Raters would not share the interpretations of the Generators (situation 1) is small. Secondly, in this research a Feature Change is considered not valid when there is even a small hint towards a difference between sets of descriptions ($p < 0.1$).

Conclusion

For the validity of images to be used in further research three criteria need to be fulfilled. First, changes should not be confounded with other factors which might facilitate the detection of Gist Changes more than the detection of Feature Changes. Second, a Gist Change should be perceived as a change affecting the gist. Third, a Feature Change should not affect the gist as it is perceived.

To guard against possible confounding factors both Gist Changes and Feature Changes are applied to the same image. Furthermore, the Gist Changes affected in general a much smaller portion of the image than the Feature Changes. Finally, the positions in the image at which the change was applied, and the types of changes were varied between images. Whether or not the changes affected the gist of an image was determined by applying the method from chapter 3. Based on the criteria it was concluded that 10 test sets are suitable for testing the hypothesis that people are more sensitive to changes affecting the gist than to changes that do not affect the gist. Experiments testing this hypothesis are reported in the next chapter.

5

Measuring change detection⁶

Imagine you are in a bar with some friends. One friend tells you that his twin brother might be joining you this evening. While you are at the bar to get some drinks, the twin brother arrives. Your friends decide to test you on whether or not you are able to see the difference between the twin brothers. Your friend goes to the restroom and his brother takes his place. In which of the following two situations are you more likely to notice that the person pretending to be your friend is actually his brother?

- Situation 1: The twin brother is dressed completely different than your friend.
- Situation 2: The twin brother is dressed exactly the same, but has one black eye.

In situation 1, noticing the switch depends on whether or not you remember what clothes your friend was wearing. Assuming you are not particularly interested in fashion, clothes can be seen to represent a feature. In situation 2, noticing the switch depends on whether or not you see the black eye. In general a black eye is meaningful (i.e., part of the gist) as it is usually a result of violence or clumsiness. In situation 2 noticing the switch is comparable to detecting a Feature Change, and in situation 1 noticing the switch is comparable to detecting a Gist Change.

In this chapter two experiments are reported, which were conducted to assess whether changes affecting the gist are better detected than other changes. First, evidence from other studies on sensitivity for semantic changes is discussed. Then the sensitivity to gist changes, compared to sensitivity to other changes is tested in two experiments. The first experiment tests whether gist changes are better detected than feature changes. The second experiment tests whether gist information is automatically retained in memory.

⁶ Research previously presented at the 27th European Conference on Visual Perception (ECVP, 2004) and the 13th Annual Workshop on Object Perception, Attention, and Memory (OPAM, 2005).

5.1 Evidence for importance of semantic meaning

It has long been suggested that even though a lot of changes can go unnoticed (see Simons and Ambinder, 2005 for an overview), changes affecting the gist will be detected (Friedman, 1979, Simons & Levin, 1997). However, direct evidence that gist is sufficient for change detection is lacking. Change detection in images and in real world scenes has been tested extensively. Besides the influence of physical changes (e.g., color, size) several studies have been reported on changes to semantic informative elements, such as change detection for objects or areas of central interest and marginal interest (Rensink et al., 1997), semantic and non-semantic changes (Werner and Thies, 2000), and consistent and inconsistent objects (Hollingworth and Henderson, 2000). The importance of semantic meaning has been demonstrated with a variety of paradigms and test stimuli. Three lines of research are discussed below.

A first line of evidence for sensitivity to semantic meaning concerns detection of changes made during a visual disruption. One paradigm often used to test sensitivity for changes made during a visual disruption is the Flicker Paradigm (Rensink et al., 1997). In a typical task in a flicker paradigm a picture (A) is followed by a mask, which is followed by an altered version of the picture (A'). The sequence is repeated until the participant indicates seeing a difference. Using a flicker paradigm, Rensink et al. demonstrated that changes to objects or areas of central interest are detected faster than changes to objects or areas of marginal interest. Note that a change to an object of central interest does not necessarily change the interpretation, and therefore the gist of an image (see also chapter 4). Another study examining change detection of changes applied during a visual disruption is a study by Werner & Thies (2000). They demonstrated that experts on the domain of American football were better at detecting football related semantic informative changes than novices. However, semantic informative changes in traffic related pictures were not detected better than non-semantic informative changes.

A second line of evidence for sensitivity to semantic meaning concerns detection of differences between images in a Memory Paradigm. In a typical memory paradigm an image is studied and memory for the image properties is tested. Either memory is tested by asking questions concerning the studied image (see for instance Melcher, 2006) or a test image is presented for which a participant has to respond whether or not the image is the same as a studied image (see for instance Friedman, 1979). Better detection of changes to objects inconsistent with the high-level category of an image has been demonstrated in memory experiments (Hollingworth and Henderson, 2000, Friedman, 1979; Biederman, 1981). Further, when an object is substituted by an object of the same conceptual category (token change) detecting the change is more difficult than when an object from a different conceptual category (type change) is substituted (Nelson, Reed, and Walling, 1976, Mandler and Johnson, 1976, Mandler and Ritchey, 1977). Also, it has been demonstrated that the level of expertise or personal interest of a participant can have an effect on memory for visual information. For instance, experienced chess players have better memory for meaningful chess position than novice players (De Groot, 1965).

A third line of evidence for sensitivity to semantic meaning concerns research revealing false recognition. Like in a Memory paradigm images are studied and memory is tested. Typical for research revealing false recognition is that the focus is less on memory for visual details than

in a difference detection memory tasks. Koutstaal & Schacter (1997) tested false recognition for semantic meaning by presenting several images that were thematically related (e.g., boats)⁷. They found that especially older participants were likely to falsely recognize thematically related new images. Potter, Staub, and O'Conner (2004) tested the amount of information retained from briefly presented images. They found that thematically related images were more likely to be falsely recognized as being old than images that are thematically unrelated. However, on average, participants did remember enough visual information from the studied images to distinguish between old and new images. Potter et al. concluded that participants did retain more information than only the abstract high level meaning of the image. This conclusion is shared by Homa and Viera (1988). They demonstrated that even after eight weeks participants retained enough information from studied images to correctly remember whether they had studied a line drawing or a thematically similar picture. On the other hand, Intraub and Hoffman (1992) demonstrated that participants could not always correctly identify whether they had seen a picture or merely read a paragraph about a picture. This indicates that the retained information is not necessarily visual, but might be conceptual. The findings by Intraub and Hoffman (1992) confirm the expectation that people remember the gist of an image (e.g., an interpretation of what the image is about).

Besides evidence for sensitivity to meaning, there are three other factors which might facilitate the detection of gist changes. Conceptual information is extracted very fast, almost automatically, and is one of the first properties encoded in memory. It is widely accepted that conceptual information (e.g., gist) from a stimulus is available very quickly (Potter, 1993). For instance, Potter (1975, 1976) demonstrated that people are able to detect a target image (specified only by a short title capturing the meaning of the image) when images are presented in rapid succession (e.g., 125 ms per image). However, memory tests for the rapidly-presented images at the end of a sequence demonstrated that the information was not stored in memory. Initially it was suggested that the extraction of conceptual information might be automatic. The automatic extraction of the conceptual information from a new image would result in disrupted encoding of a previous image (Potter, 1976). The disrupted encoding would explain why memory for briefly presented images is limited whereas the image recognition performance is high. Further research revealed that the degree of conceptual encoding in memory depends on how soon the encoding process is disrupted (Intraub, 1981, 1984). Intraub (1984) demonstrated that while processing of new conceptual information is not necessarily automatic, ignoring new information is difficult. A study by Tatler, Gilchrist, & Rusted (2003) demonstrated that not only is gist perceived very quickly, it is also one of the first properties from an image stored in memory. By systematically varying the presentation time of images, they found that the gist of an image is encoded in memory within 1 second. Increasing the presentation time did not lead to better performance for gist related questions. In contrast, performance for other properties (such as shape, color, position, and relative distance) did benefit from longer presentation times.

The two experiments reported in this chapter were designed to examine the role of gist in detecting a difference between two images based on memory. In both experiments images

⁷ With their research they extended the Deese/ Roediger-McDermott paradigm to visual stimuli. Deese/ Roediger-McDermott paradigm: examines memory illusions in a standard list-learning procedure, where words related to grouped words are falsely recalled.

were presented for the duration of five seconds in a study phase. The duration of five seconds gives participants enough time to extract the gist, and encode feature information (Tatler et al., 2003). In a test phase, altered versions of the studied images are presented, to assess what kind of information from the studied images participants do compare during a difference detection task. Whether or not a difference between the studied image and the image in the test phase is detected depends on the retained information and the information which is compared. The first experiment tests whether or not a pictorially relative small change affecting the gist of the image will be detected in an image otherwise visually very similar to the studied image, and compares sensitivity for gist changes to sensitivity for pictorially larger changes not affecting the gist. The second experiment sets out to test whether participants encode gist information in memory automatically. That is, whether participants remember image properties without specific instructions preparing them for a memory test.

5.2 Experiment 1: Intentional encoding

As the gist of an image is so quickly perceived and encoded in memory, we hypothesized that gist is one of the first aspects used during a difference detection task. To test this hypothesis, two types of changes were made to images (see also chapter 4): a change that affected the gist of the image, and a change that did not affect the gist. If people are more sensitive to gist than to other image properties, participants should notice a change affecting the gist more often than a change not affecting the gist. Also, Experiment 1 examines whether or not the quickly available gist facilitates quick detection of changes affecting the gist.

Method

Participants. Twenty-one volunteers (5 female) from the Technische Universiteit Eindhoven participated in the experiment. They received 4 euros for their participation. The participants had no knowledge of the hypothesis to be tested and did not participate in any earlier study using similar stimuli.

Materials. Stimuli were 18 test sets; 10 "critical" sets and 8 filler sets⁸. A critical test set consists of a Target image (Original image), a Feature Change, a Gist Change image and two Distractor images (Figure 5.1). Target images were clipart images portraying a variety of in and out-door scenes collected from Microsoft Clipart. The Feature Change and the Gist Change images were altered versions of the Target image. They were pictorially similar to the Target image except for the altered object or area. For the Feature Change, objects were added, deleted, mirrored, substituted or changed in color in such a way that the gist of the image is not affected. On average the changed aspects for Feature Change images affected 10% of the pixels from the original image. For the Gist Change, objects were added, deleted, mirrored, and substituted in such a way that the gist of the image is affected. On average the changed aspects for Gist Change images affected 2% of the pixels from the original image. The Gist Change and the Feature Change images were validated in an independent study discussed in chapter 4. For each Target image two pictorially different clipart images were selected to serve as Distractor⁹ images. The eight filler sets were initially intended as test set.

⁸ Critical test sets were test sets 1-10 in Figure 4.2. Test sets 11-18 in Figure 4.2 served as filler test sets.

⁹ Initially there were two types of distractor images, one with a gist similar to the target image, and the other one with a different gist. The definition of gist used for the selection of the images was: high level category. The

The filler sets also contained a target image, two types of decoy images, and two distractors. The validation method as described in chapter 4 showed that the intended gist changes and/or the intended feature changes were not valid. The results for the filler sets are not reported here.

The experiment was programmed in E-prime (Psychology Software Tools, 2002). All images were 300 pixels high or wide depending on the largest dimension. They were presented in color and centered on a 17-inch monitor with a screen resolution of 1024x768 pixels. Participants were seated approximately 0.5 meter from the computer screen.

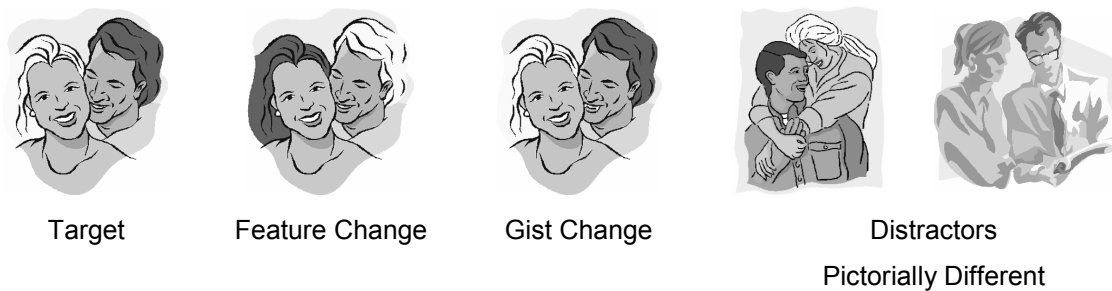


Figure 5.1. Example of a "critical" test set consisting of a Target, Feature Change (swapped hair color), Gist Change (added fangs), and two pictorially different Distractor images.

Design and Procedure. The performance (accuracy and response latency) was tested on four types of images (Target, Feature Change, Gist Change and Distractors) within subjects in an Old/New memory paradigm. The experiment consisted of a study phase and a test phase. In the study phase the 18 Target images were displayed sequentially on a white background for five seconds, followed by a black screen for five seconds. The presentation order of the Target images was fixed. In the test phase, both old (Target) and new (Feature Change, Gist Change and Distractors) images were presented, for which participants had to respond either "Same" or "Different" (from here the task will be referred to as Same/Different task¹⁰). The test phase contained 108 trials. For each studied Target image there were six trials in the test phase. Each Target image appeared twice in the test phase, the Feature Change, Gist Change and the two Distractor images each appeared once. The trials were arranged in six blocks, each block contained one trial for each of the 18 Target images. The order of the trials in each block matched the order of the Target images in the study phase. The order of the type of image on a trial (Target, Feature Change, Gist Change or Distractor) was randomized between

validation procedure as described in chapter 4 revealed that, except for four images, all pictorially different images had a different gist than the target image. Therefore, the initial hypothesis that different images with a similar gist would be more likely to be confused with the target image than different images with a different gist could not be tested. Although some of the distractor images are thematically more related to the target image than others, there was no difference in performance for the two types of distractors.

¹⁰ Strictly, a Same/Different task refers to a task where participants need to indicate whether or not two presented stimuli are the same. In this thesis the name Same/Different is used to indicate an Old/New task as the words Same and Different more closely reflect the decision task in the experiments than do the words Old and New.

participants. For each trial accuracy and response latency were recorded. Response latency was measured from the time the test image was presented until a response.

Participants were tested individually. They were informed that the experiment was designed to gain insight in the speed of difference detection between images. They were instructed that in the test phase they would be asked to judge whether an image was **exactly** the same as the image from the study phase, or different. Participants were instructed to press a green key (left, "Z" key) to respond "Same" and a red key (right, "/" key) to respond "Different". They were urged to respond as quickly and accurately as possible. They were informed that the time limit for responding was five seconds. Each new trial started with a fixation cross in the middle of the screen. Participants were not informed on the types of changes that were applied to the images.

To familiarize participants with the procedure, the available time to study an image, and the kind of test images they could expect in the experiment, the experiment started with a practice trial. The practice trial consisted of one image in a study phase and six corresponding images in the test phase (Target image twice, Feature Change, Gist Change and two Distractors). In the practice test trials participants received feedback on the accuracy of their response and their response time. If they did not respond within the appointed five seconds they were urged to respond faster the next time. After the practice trial the actual experiment started with the study phase. No feedback was provided during the experiment. The test phase started approximately 2 minutes after the study phase.

Results

For each participant the performance (accuracy and response latency) for the four types of images (Target, Feature Change, Gist Change and Distractors) was calculated by combining the responses for the 10 critical test sets. For Target images and Distractor images the mean accuracy for each participant is based on 20 trials. For Feature Change images and Gist Change images, the mean accuracy is based on 10 trials. Mean reaction times are only based on correct responses. When a participant failed to respond within five seconds, the trial was disregarded (1% of all trials). Only the data obtained for the critical test sets are reported.

Figure 5.2(a) displays the mean proportion correct "Different" responses for Feature Change, Gist Change and Distractor images. For Target images Figure 5.2(a) also displays the mean proportion incorrect "Different" responses. The mean response latency for correct responses for each type of image is displayed in Figure 5.2(b). For each mean score in Figures 5.2(a) and (b) the 95 % confidence interval is shown. The 95% confidence intervals indicate an estimate of the mean score for the population. They directly provide insight in how likely it is that the same patterns will occur when the experiment is repeated. When there is no overlap between two confidence intervals this indicates that the two corresponding means are significantly different ($p < 0.05$). Considering the size of the effects most statistical tests are superfluous. Statistical tests are only provided for the effect of interest (Feature Change vs. Gist Change). For an overview of the rationale behind visual data interpretation see Loftus and Masson (1994).

Gist Changes were detected significantly more often ($t(20)=8.82, p < 0.001, r=0.89^{11}$) and faster ($t(20)= 3.14, p = 0.005, r=0.57$) than Feature Changes (Figure 5.2(a) and (b)). Further, the mean accuracy for Distractor images indicates that participants were very unlikely to confuse the pictorially different Distractor images for the Target image (Figure 5.2(a)). Furthermore, responses for Distractor images were significantly faster than accurate responses for all other types of images (Figure 5.2(b)).

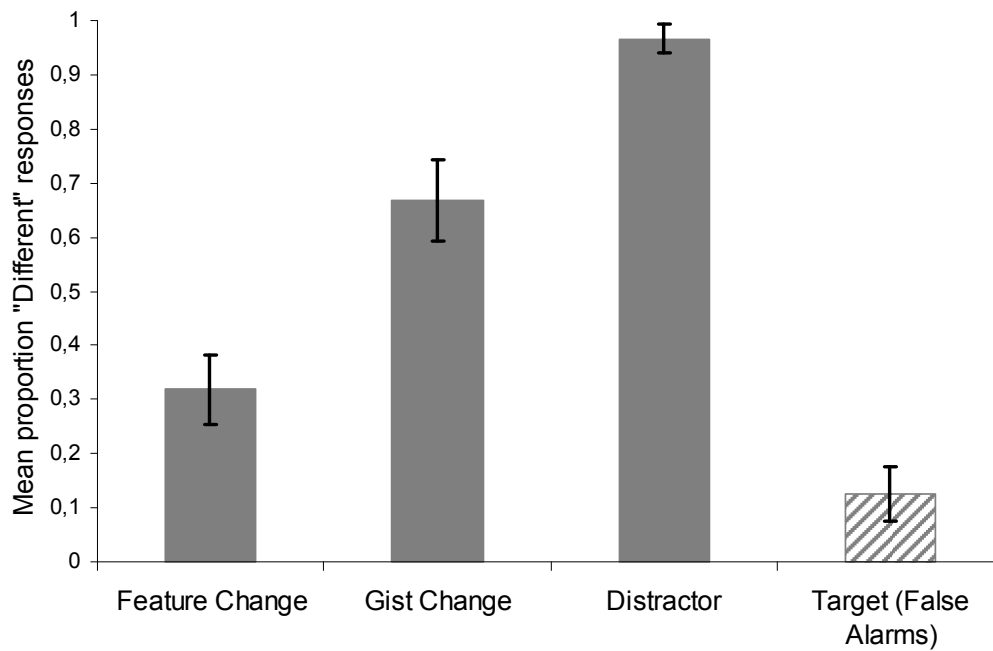
Discussion

Experiment 1 was designed to test whether or not Gist Changes would be easier to detect than Feature Changes. The results indicated that pictorially small Gist Changes were indeed more visible than the pictorially larger Feature Changes. The large difference in detection accuracy between Gist Changes and Feature Changes clearly suggests participants were more sensitive to conceptual changes than visual changes. The significantly faster responses for Gist Change images, compared to responses for Feature Change images, indicate that gist might be the first property compared during a difference detection task. The relatively fast detection of Gist Changes compared to Feature Changes is consistent with the fast detection of changes to semantically informative objects in a Flicker Paradigm (Rensink et al., 1997; Werner and Thies, 2000). The fast and accurate responses to the Distractor images indicate that visual similarity of the image as a whole was the first property used for recognition. The limited false recognition of the Distractor images is consistent with the results from Potter et al. (2004) and Homa and Viera (1988), who suggested that people encode more than the abstract meaning of an image.

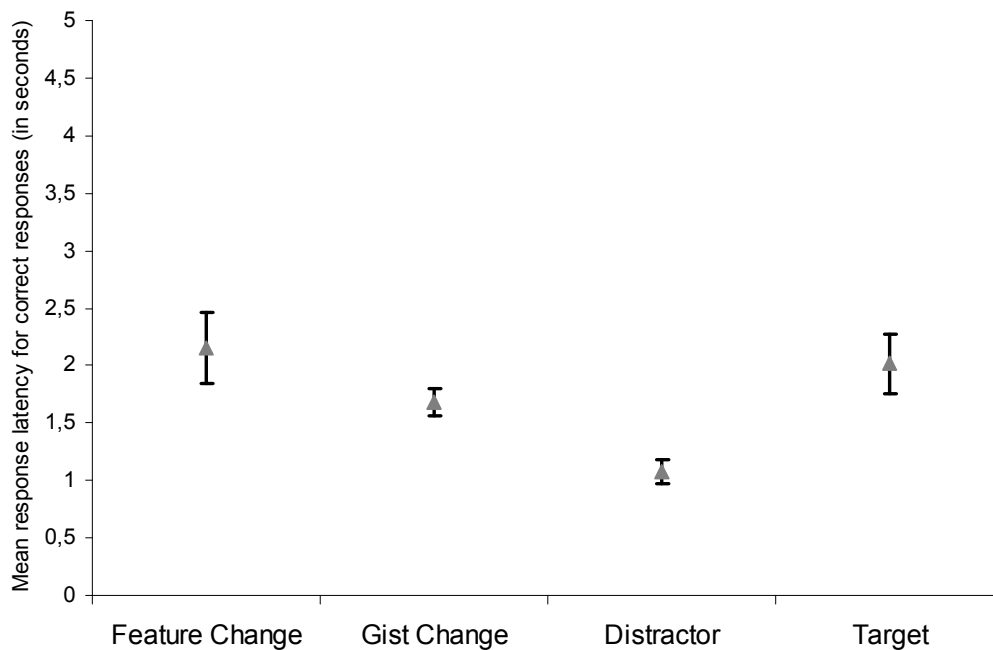
The repeated measure design (i.e., each participant saw all types of test images for each Target image) could have affected the results in three ways. First, the design could have resulted in an overestimation of the performance for the Distractor images. When a participant would have initially not retained much information from the Target image in the study phase, encountering the Target image or the visually similar Gist and Feature Change images in the test phase, it could have served as a reminder. If participants initially do not remember much, false recognition of the Distractor images will be higher in Experiment 2, when for each Target only one image is present in the test phase. Analysis of the trial in which the Distractor image was the first image tested from a test set revealed a very small indication of an effect of the repeated measure design for Distractor images. The mean accuracy for Distractor images on the first trials was 95% whereas the mean accuracy for distractor images for the remaining trial was 97%.

¹¹ The r is a measure for Effect Size and is calculated from the test statistic (t) and the degrees of freedom (df). The r can vary between 0 and 1. Following Cohen (1988) a $r = 0.10$ indicates a small effect, $r = 0.30$ indicates a medium effect and $r = 0.50$ indicates a large effect.

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$



(a)



(b)

Figure 5.2. Results for Experiment 1.

(a) Mean proportion correct "Different" responses for Feature Change, Gist Change and Distractor images and mean proportion incorrect "Different" responses (False Alarms) for the Target image. Error bars represent the 95% confidence interval for the mean.

(b) Mean response latency for correct responses for Feature Change, Gist Change, Distractor and Target images. Error bars represent the 95% confidence interval for the mean.

A second way in which the repeated measure design could have affected the results was that participants got confused. Some participants reported after the experiment that they had sometimes based their decision on a previous image from the test phase instead of on the studied image. For instance, one participant explained that he had responded "Different" for a Target image, because the colors had been different from a previous image (Feature Change) for which he had responded "Same". When the Target image was presented for the second time in the test phase, it occurred to the participant that this was in fact the same as the studied image. This kind of confusion may have affected the responses for Target image, Feature Change images and Gist Change images.

A third way in which the repeated measure design could have affected the results is that the repeated presentation of the images might have caused learning. Analyses of the mean accuracy for Target images, Feature Change and Gist Change images for the first trial on each test set compared to the mean accuracy for the remaining trials in the test phase reveals a clear disadvantage for the changed images compared to the Target image. On the first trials participants seemed more likely to respond same to all images similar to the studied image. This resulted in relatively high accuracy for Target images in the first trials compared to the later trials (95% vs. 86%). For Feature Change images responses were less accurate during the initial trials compared to the mean accuracy during the remaining trials (21% vs. 34%). This effect was even more pronounced for Gist Change images with a mean accuracy of 38% during the first trials and a mean accuracy of 74% for the remaining trials. Presumably, participants used different strategy during the first and the remaining trials. However, it is also possible that the better performance on the later trial was a result of refreshed memory for the images (as discussed in relation to the Distractor images above). To prevent unwanted learning effects, in Experiment 2 participants were only tested on one image for each studied image.

5.3 Experiment 2: Incidental encoding.

Experiment 1 demonstrated that participants were more sensitive to changes affecting the gist of an image than to other changes. It also demonstrated that a relatively small difference affecting the gist of an image is sufficient for change detection, whereas relatively larger changes not affecting the gist are not necessarily detected. For an indication of the role of gist in real life situations it is useful to examine whether or not the gist of an image is processed automatically and retained in memory without specific instructions to do so. In real life situations one never knows what information might be necessary or worthwhile to remember.

Experiment 2 was designed to test whether the gist of an image is encoded during an incidental study task, and whether gist is remembered over a longer period of time (compared to the retention period in Experiment 1). Also, Experiment 2 sets out to remedy the repeated measure design from Experiment 1. In Experiment 2 the four types of images from each test set are tested between subjects. So for each studied Target image only one type of image (Target, Feature Change, Gist Change or Distractor) was present in the test phase. By presenting only one image from each test set in the test phase, the responses will more directly reflect what is remembered from images than the responses from Experiment 1. If people do not remember much of what they have seen, performance for Distractor images should be less accurate than in Experiment 1.

Method

Participants. Participants were 163 students from Tilburg University (92 female, mean age 21.1). The experiment was part of a computerized questionnaire containing several unrelated decision tasks. Upon completion participants received 5 euros payment. The data from five participants were disregarded because these participants indicated that they suspected a memory test. Consequently, data from these participants could not be treated as data from an incidental encoding task.

Stimuli. Stimuli were a subset of the test sets used in Experiment 1; six "critical" sets and four filler sets¹². The critical test sets each consisted of a Target, a Feature Change, a Gist Change and two Distractor images. The mean size of the area affected by the change was smaller for Gist Change (3%) than for Feature Changes (15%).

Design and Procedure. Performance (accuracy and response latency) was tested for four types of images (Target, Feature Change, Gist Change and Distractors) within subjects in an Same/Different incidental encoding memory paradigm. In contrast to Experiment 1, for each Target image only one image was present in the test phase.

The experiment consisted of a study phase and a test phase, separated by a number of unrelated decision tasks. In the study phase, participants saw ten images. To prevent participants from consciously memorizing the images, the study phase was disguised by a cover story. Participants were asked to study pairs of images and state which of the two images they preferred. Participants were told that the purpose of the experiment was to gain insight in the relationship between self-reported creativity and image preference. They first rated their creativity and talent for drawing on an eight-point scale. Next, an image was presented for five seconds, followed by a second image. A blank screen (2 seconds) separated the images. After the second image participants were asked to press a button (by means of a mouse click) indicating a preference for either the first or the second image. This procedure was repeated five times, until all ten target images had been displayed. Several unrelated tasks followed before the test phase of the experiment started.

In the test phase the memory for the studied images was tested in ten trials, one for each studied image. In the test phase participants were asked to indicate as accurately and quickly as possible whether an image was exactly the same as one of the images they had seen in the image preference task, i.e., the study phase. For each image from the study phase, the image in the test phase could be either the Target, or Feature Change, or Gist Change or a Distractor image. There was no time limit for responding.

The test phase started with a practice trial. Participants studied one image and were tested with a Feature Change image. After responding to the practice trial participants were informed that the test image had been different. The actual test phase started after the practice trial. No feedback was provided during the experiment. The test phase started approximately 15 minutes after the study phase. To ensure the memory for the studied images were a result of incidental encoding, upon completion of the test phase participants were asked whether they had expected a memory test. When a participant indicated a suspicion, his or her data were disregarded.

¹² The critical test sets were sets 1, 2, 5, 6, 9 and 10 in Figure 4.2. Test sets 15-18 served as filler sets.

Results

To reduce the effect of extremely long response latencies, response latencies (and the corresponding accuracy scores) larger than three standard deviations from the overall mean, were disregarded (2% of all critical trials). Only the data for the six critical test sets are reported. For each participant performance (accuracy and response latency) for the four types of images (Target, Feature Change, Gist Change and Distractors) was calculated by combining the responses for the six critical test sets. The resulting accuracy scores for each of the four types of images were based on responses for one, two or three trials. Mean response latencies were only based on correct responses.

Figure 5.3(a) displays the mean proportion correct "Different" responses for Feature Change, Gist Change and Distractor images. For Target images Figure 5.3(a) displays the mean proportion incorrect "Different" responses. The mean response latency for correct responses for each type of image are displayed in Figure 5.3(b). For each mean score in Figures 5.3(a) and (b) the 95 % confidence interval is shown.

As in Experiment 1, Gist Changes were detected more often than Feature Changes ($t(97)=7.68$, $p < 0.001$, $r=0.61$). In contrast to Experiment 1, there was no difference between the latency for detecting Gist and Feature Changes ($t(13)=1.58$, $p > 0.10$, $r=0.40$). Note however that the mean latency for Feature Changes is based on a limited amount of data.

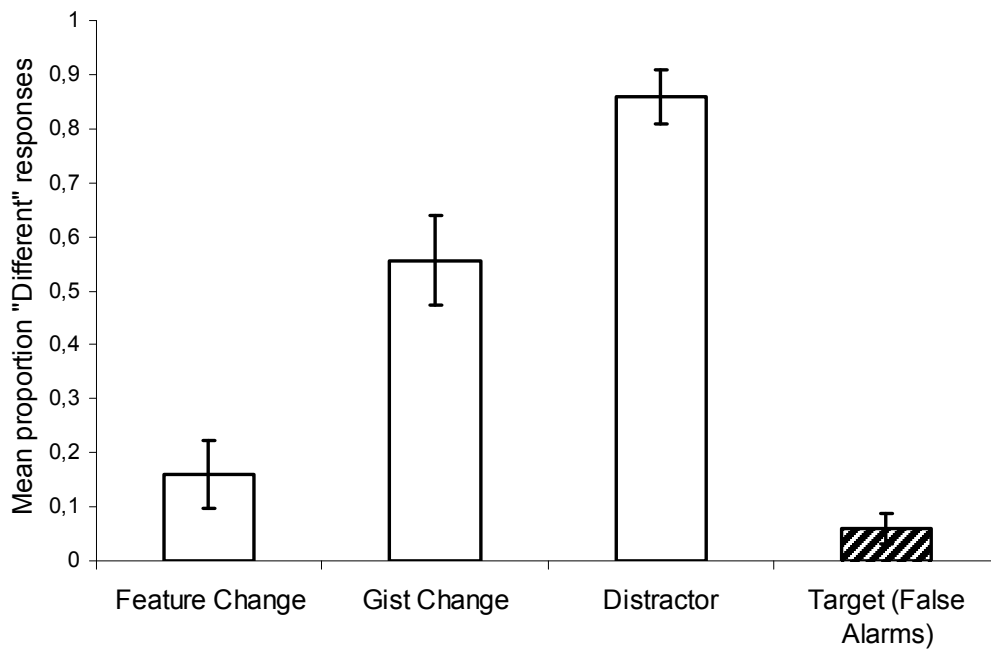
Discussion

Experiment 2 was designed to test the effect of incidental encoding on gist and feature information. The significantly better detection of Gist Changes compared to Feature Changes indicates that the gist of an image was extracted and retained in memory.

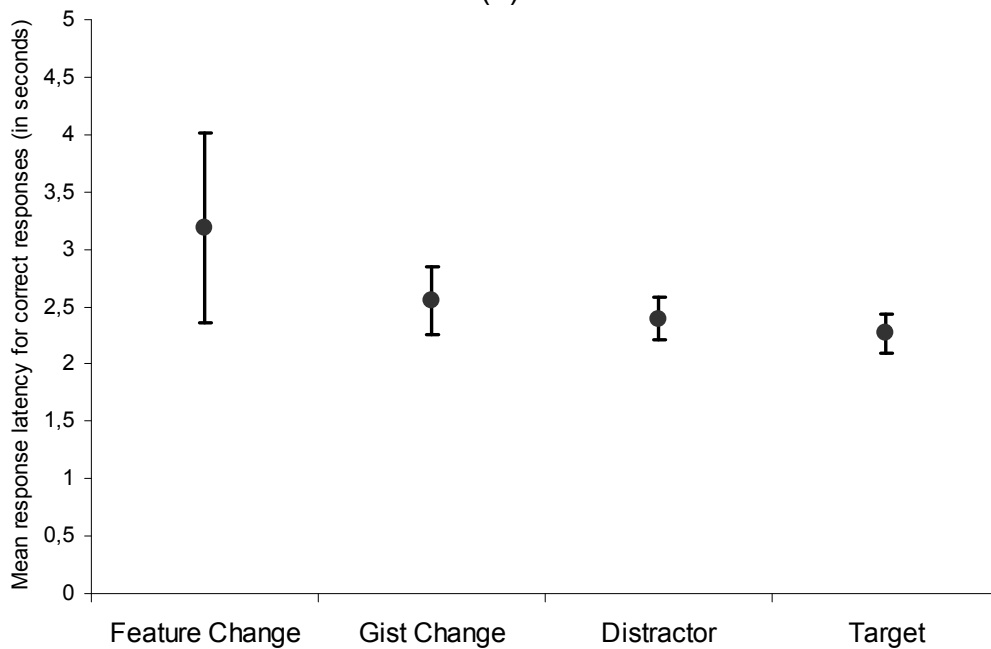
Further, Experiment 2 was designed to remedy possible confounds from the repeated measure design used in Experiment 1. The high mean proportion correct responses for Distractor images (0.86) indicates that participants did remember enough from the studied images to correctly identify the Distractor images as being "Different". It can therefore be concluded that the high accuracy for Distractor images in Experiment 1 can not be simply a result of the repeated measure design. This implies that people do indeed retain more information from an image than just the central theme of the image.

In contrast to Experiment 1, there was no significant effect of the type of image on the speed of difference detection. For all types of images the responses were slower in Experiment 2 than in Experiment 1. Partly this should be attributed to the use of a mouse to select the response button, whereas in Experiment 1 participants responded by pressing a red or green key. Another factor that might have contributed to slower responses in Experiment 2 is that there was no time limit for responding. Perhaps the time limit set in Experiment 1 encouraged participants to be fast, whereas in Experiment 2 participants might have focused on being accurate.

Another interesting question is whether or not the incidental coding task affected memory for features more than memory for gist. This question is addressed in the next section.



(a)



(b)

Figure 5.3. Results for Experiment 2.

(a) Mean proportion correct "Different" responses for Feature Change, Gist Change and Distractor images and mean proportion incorrect "Different" responses (False Alarms) for the Target image. Error bars represent the 95% confidence interval for the mean.

(b) Mean latency for correct responses for Feature Change, Gist Change, Distractor and Target images. Error bars represent the 95% confidence interval for the mean.

5.4 Comparing results from Experiment 1 and Experiment 2

According to Potter (1976) and Intraub (1984) the gist of an image might be processed almost automatically. Combined with frequent claims that people only retain an abstract representation of an image (see, for instance, Simons & Levin, 1997), this leads to the hypothesis that features are less likely to be stored in memory than gist when people do not look at an image with the purpose of remembering. If features are less likely to be remembered than gist, the detection of Feature Changes in Experiment 2 should be affected more than the detection of Gist Changes in Experiment 2, compared to the detection measured in Experiment 1. On the other hand, recent research revealed that incidental encoding does not greatly affect memory for features (Castelhano & Henderson, 2005). If this is the case, detection of Feature Changes and Gist Changes might be equally effected by the incidental encoding task.

Unfortunately the difference in the design of the experiments does not permit a direct comparison of the results from Experiment 1 and Experiment 2. However, d' scores¹³ can provide an indication for differences in sensitivity for Feature Changes, Gist Changes and Distractors between both experiments. The d' scores for both experiments are presented in Table 5.1. The proportion of Hits is calculated by the sum of all correct "Same" responses for all Target images divided by all Target trials (excluding the trials with late responses). For the critical test sets, Feature Change, Gist Change and Distractor images the proportion of False Alarms is calculated by the sum of all incorrect "Same" responses divided by the sum of all corresponding trials (excluding the trials with late responses).

Further, a bias measure¹⁴ (β) is provided in Table 5.1. The bias (β) indicates whether participants were more inclined to respond "Same" or "Different". For all types of images the sensitivity (d') is lower in Experiment 2 than in Experiment 1. Sensitivity for Feature Changes seems to decrease more (from $d'= 0.75$ in Experiment 1 to $d'= 0.49$ in Experiment 2) than sensitivity for Gist Changes (from $d'= 1.67$ in Experiment 1 to $d'= 1.60$ in Experiment 2). This suggests that detection of Feature Changes suffers more from an unsuspected memory test than detection of Gist Changes. However, it is also possible that the suggested decrease in Feature Change detection is not a result of the unexpected test, but a result of the relatively long retention period between study and test phase in Experiment 2 compared to Experiment 1.

¹³ $d' = |z(\text{Hits}) - z(\text{False Alarm})|$

¹⁴ The likelihood ratio bias measure (β) can be calculated by the formula: $\beta = e^{cd'}$, where c is the criterion location ($c = -0.5 * |z(\text{Hits}) + z(\text{False Alarm})|$). See Macmillan and Creelman (2005).

Table 5.1. Sensitivity (d') scores and likelihood ratio bias β (in parentheses) for each of the three types of images tested in Experiment 1 and 2.*

	Feature Change		Gist Change		Distractor	
Experiment 1						
Intentional encoding	0.75	(0.54)	1.67	(0.54)	3.09	(2.81)**
Experiment 2						
Incidental encoding	0.49	(0.53)	1.60	(0.31)	2.83	(0.72)

* β can vary between 0 and ∞ , a value between 0 and 1 indicates bias toward responding "Same", a value of 1 indicates no bias, and a value > 1 indicates bias to responding different.

** The relatively large number of "Different" responses for the Target image and the small number of "Same" responses for the Distractor images led to the large bias for responding "Different".

General discussion and conclusions

The objective of this chapter was to examine whether changes affecting the gist of an image are better detected than other changes. From the results of Experiment 1 it can be concluded that the gist of an image is remembered and gist information is used during a change detection task. Performance for changes affecting the gist of an image was faster and more accurate than performance for other changes. Experiment 2 demonstrated that the gist of an image is extracted and remembered without specific instructions to do so. Comparison of the data from Experiments 1 and 2 suggests that gist is more likely to be remembered, without specific instructions to prepare for a memory test, than features. In short: this research clearly indicates that gist plays a large role in difference detection. If a change affects the gist of an image, the change is more likely to be detected than if it does not affect the gist. So when your friends try to fool you by substituting one friend by his twin brother, you are more likely to notice the switch if something meaningful (black eye) is different than when they are dressed completely different.

The outcome of the two experiments reported in this chapter is totally consistent with the general assumption that changes affecting the gist are detected (e.g., Simons & Levin, 1997). The fact that participants were unlikely to falsely recognize the visually different thematically related Distractor images is consistent with the findings of Potter et al. (2004) and Homa and Viera (1988).

Although the results presented in this chapter suggest that features are to a minor degree used during a difference detection task, this does not necessarily mean that feature information is not stored in memory. Recent research has revealed that more is remembered than is compared (e.g., Mitroff, Simons, & Levin, 2004). In the next chapter the question is addressed whether feature information is retained, but is not used during a difference detection task.

6

More than you compare¹⁵

Imagine you are in a bar with a friend. After a visit to the restroom you return to your friend and find that suddenly there are two. The friend happens to have an identical twin. The only visible difference between your friend and his twin are their clothes. Would you be able to recognize your friend when the twins are standing side by side?

For successful recognition of your friend, information about his clothes needs to be stored and retained in memory and needs to be compared to the information available in front of you. The experiments in chapter 5 demonstrated that participants were not very likely to detect a change in features (e.g., clothes). Although traditionally the inability to detect changes was interpreted as evidence for limited memory of images, limited memory is not necessarily the cause of failure to detect changes. Scott-Brown, Baker and Orbach (2000) first pointed out the idea that an inability to detect differences might be caused by a failure to compare the information necessary to detect the difference. They demonstrated that participants found it difficult to detect difference between two stimuli (patterns composed of patches of various contrasts) presented in the central visual field within a single fixation. Recent research using more complex stimuli than Scott-Brown et al. (2000) has demonstrated that poor change detection should indeed not always be attributed to sparse memory. It is also possible that poor detection of Feature Changes is caused by a failure to compare the stored information of the object before the change with the object after the change (Mitroff, Simons & Levin, 2004; Simons, Chabris, Schnur, & Levin, 2002; Angelone, Levin, & Simons, 2003; Hollingworth, 2003). As the need for comparison is apparent when the twins stand side by side, in such a case failure to recognize your friend can only be attributed to a lack of memory for the distinguishing properties.

As the aim of this thesis is to gain a better understanding for the concept gist and its role in memory for images, it is interesting to know whether the superior detection of Gist Changes displayed in chapter 5 results from better memory for gist or, alternatively, from a better comparison of gist. As change detection depends on memory and comparison, it is possible

¹⁵ Preliminary results of the research described in this chapter were presented at the 6th annual meeting of the Vision Sciences Society (VSS, 2006).

that gist information is not remembered better than feature information, but is more likely to be compared during a difference detection task than feature information. To examine whether Gist Changes are more likely to be detected because the gist is more likely to be compared, this chapter sets out to test whether or not people do encode and store feature information. In the next section, research is reported that demonstrates that it is plausible that more information about an image is stored in memory than results from change detection experiments suggest. In an experiment it is tested whether or not a failure to compare stored information to the new situation can explain the poor Feature Change detection compared to Gist Change detection as found in chapter 5.

6.1 Previous evidence for preserved representations

There are two possible explanations for the poor performance on Feature Change detection compared to performance on Gist Change detection. Either information about features is not stored, or it is not retrieved and/or used during the change detection task. The first possibility, features are not remembered, is consistent with the *visual transience hypotheses* (term adopted from Hollingworth and Henderson, 2002). According to the *visual transience hypotheses* people have only limited memory for images. This memory contains the gist, spatial layout and a few attended objects. Other objects and regions are perceived, but quickly forgotten once attention is withdrawn (e.g., Rensink, 2000b; Wolfe, 1999). This theory is the traditional explanation for Change Blindness; it is caused by sparse internal representations of an image. People only encode and store abstract information (e.g., the gist) and some attended objects (e.g., Rensink, O'Regan, & Clark, 1997; Simons & Levin, 1997).

The second possibility is that features are stored in memory but are not compared to the features present in the new situation. This view is consistent with the *visual memory theory*, as proposed by Hollingworth and Henderson (2002). According to *visual memory theory* a relatively detailed scene representation is built up in memory across eye fixations. In contrast to the *visual transience hypotheses*, *visual memory theory* suggests that the representation of an image is stable across successive fixations and stored in Long Term Memory (LTM) even when attention is withdrawn (see also Henderson and Hollingsworth, 2003; Hollingworth, 2005). Change Blindness is explained by either a lack of attending the changed aspect during the initial viewing of the image, or a lack of attending and comparing the stored information during the test phase (Hollingworth & Henderson, 2002). Good detection of Gist Changes is predicted by both the visual transience hypotheses and the visual memory theory¹⁶.

Recently, the hypothesis that more is remembered than is compared during change detection experiments has been examined, with varying success. For instance, Levin, Simons, Angelone, & Chabris (2002) found no convincing evidence that people who did not notice a change had some memory of the situation from before the change. Their study involved an unexpected change in an incidental encoding situation during real-world interaction. The fact

¹⁶ Possibly the difference between the transience hypothesis and the visual memory hypotheses is only appropriate for short-term memory or saccadic memory. However, at a less theoretical level it is clear there is a difference between the researchers associated with the two theories in the extent of visual memory. For instance: Rensink (2000a) states that preserved representations are relatively sparse, containing less than 10% of the information, whereas Hollingworth (2003) states that preserved representations can be detailed enough for participants to detect whether an object is rotated.

that the change was unexpected and participants were not aware that the interaction was part of an experiment might explain why participants failed to encode information of the pre-change situation. However, similar studies with unexpected changes during real-world interaction and video clips reveal that people do remember aspects which would have made change detection possible (Simons, Chabris, Schnur, & Levin, 2002; Angelone, Levin, & Simons, 2003). This suggests that the information necessary to detect the change was available but not compared in the post-change situation.

Hollingworth (2003) demonstrates that also in experiments where participants knew they could expect changes, participants remembered more from an image than they used during the task. The Hollingworth experiments tested whether change detection would increase when the changed object was cued during the test phase. Participants viewed an initial scene for 20 seconds. During the test phase the (possibly) changed object was sometimes identified by means of an arrow and sometimes it was not identified. Detection of a change was better when the object was cued. This implies that at least some change-detection failures could be caused by comparison failure.

Further evidence for preserved representations comes from Mitroff, Simons, & Levin (2004). On each trial participants studied an array of objects for 1 second followed by a blank screen for 350 milliseconds. The second array (with one different object) appeared also for 1 second. Recognition was tested by means of a 2 Alternative Forced Choice (2AFC) task for the pre- and post- change object. They successfully demonstrated that participants were able to recognize both the pre- and the post- change object above change levels, even when they had not detected a change. The results indicate that information was available to the participant, based on which both pre- and post- change objects could be correctly identified. As people who did not detect a change were also able to correctly identify the post-change object above change levels, it can be concluded that the available information was not always used to detect a change.

It is not hard to imagine that people do not constantly compare every aspect stored in memory when they do not expect a change. For instance, people do not change into a different person in the middle of a conversation, so usually constant comparison is not necessary (e.g., Simons, Chabris, Schnur, & Levin, 2002). In change detection experiments on the other hand (e.g., Mitroff et al., 2004) comparison is necessary. Indeed, given the task (determine whether something is different) it might come as a surprise to find that not everything stored in memory is compared.

Two reasons for failing change detection caused by mechanisms other than limited memory are discussed below. The first reason can be found in a comparison mechanism with a threshold. The second reason can be found in a perception mechanism, in which only the first-impression is perceived. The threshold mechanism is addressed first. During a change detection task a participant needs to determine if something is different. Whether an image is considered the Same or Different depends on a decision threshold. Such a decision threshold might be time-related or strength-related. With a time-related threshold, an image is considered Same when after a certain amount of comparison no difference is detected. The amount of time spent on comparing is set by the participant himself, and might differ between people. It is possible that during the comparison of the stored information with the changed situation people follow a hierarchical path. As gist is perceived very quickly, gist might be the

first aspect compared. When no difference is detected during the comparison of the gist, only then the features are compared. When no difference is detected within the time limit set by the participant, an image is considered to be the Same. The hierarchical path (gist compared first, features compared later) might then explain why Feature Changes are less likely to be detected than Gist Changes. Considering that in the experiments in chapter 5 participants were urged to respond quickly and accurately, it is possible that the poor feature change detection is partly caused by time limited comparison. The hierarchical path of comparison (gist compared first, features compared later) can be regarded as a top-down process. High-level information is compared first and lower-level (more perceptual) information is compared later. Such a top-down process during comparison is consistent with the Reverse Hierarchy Theory proposed by Ahissar and Hochstein (2004).

It is also possible that the decision threshold is strength related. A strength-related threshold might work in two ways. First, a participant might decide there is something different only when something important has changed. The strength-related decision threshold terminates the comparison process when the important properties have been compared. As gist is more important than features, such a strength-related threshold would favor the detection of gist changes. The second way a strength-related decision threshold might work is that the threshold is based on expectations on the visibility of a change. A participant might only decide there is something different when the difference is clearly visible. What kinds of changes are classified as clearly visible can depend on the visibility of other changes encountered in an experiment. The classification of the visibility of a change is relative to other changes. As gist changes are relatively easy to detect, a participant could have assumed that all changes would be easy to detect. Again, a decision threshold based on expectations on the visibility of differences would explain why Feature Changes are less likely to be detected than are Gist Changes in a change detection task. Both the time-related and strength-related decision thresholds explain why stored information from the pre-change situation is not compared with the post-change situation. The decision threshold marks the moment at which the comparison process stops.

The second reason why not all information from an image stored in memory would be compared to a post-change image can be found in a perception mechanism. Features from the pre-change situation are encoded because they are used to interpret the image, but not all of those features need to be re-attended in the post-change situation (Friedman, 1979; DiGirolamo & Hintzman, 1997). As the primary goal of perception is to understand meaning, features are not relevant once an image is interpreted. During the post-change situation only the gist is compared. As long as the gist is the same, the whole image is considered the same. The details (e.g., features) are filled in from memory. See Simons (2000) for a more detailed discussion of this first-impression explanation of Change Blindness. DiGirolamo & Hintzman (1997) found strong evidence that people do encode more details from a first image than from a second image. However, the first-impression explanation can not explain all accounts of failed comparison. In the Mitroff et al. (2004) studies, participants were able to recognize both pre-change and post-change objects better than predicted by chance alone. This indicates that, in contradiction to a first-impression explanation, participants did indeed perceive information from the post-change objects. In summary, it is possible that the poor detection of Feature Changes is due to a failure to compare the stored information from the pre-change situation to the post-change situation, instead of poor memory for feature information.

To examine if this is the case an experiment was conducted in which results from a Same/Different difference detection test are compared directly to the results of a 4 Alternative Forced Choice (4AFC) recognition test. In a forced choice task participants are able to focus on the differences between images, and compare the differing properties to the information retained in memory. If changes are not detected more often in the 4AFC task than in the Same/Different task, this suggests that the necessary information was not retained in memory. If, on the other hand, change detection performance is better in the 4AFC task than in the Same/Different task, this implies that indeed more is remembered than is compared.

6.2 Experiment 3: a Same/Different task and a 4AFC task.

The experiment reported in this section should provide answers to two questions. First, it will be examined whether the poor detection of Feature Changes compared to the detection of Gist Changes as demonstrated in the experiments reported in chapter 5 are likely to be caused by a failure to compare features during a change detection task. If this comparison failure is part of the reason why Feature Changes are more difficult to detect than Gist Changes in a change detection task, sensitivity to Feature Changes in a forced choice task should be better than in a Same/Different task. Second, to assess whether comparison failure is the sole reason for poor detection of Feature Changes, the sensitivity to Feature Changes is compared to the sensitivity to Gist Changes. If the only reason why Feature Changes are so poorly detected is the fact that features are less likely to be compared during a change detection task, then sensitivity to Feature and Gist Changes should be equally good in a forced comparison task. On the other hand, it is possible that the poor detection of Feature Changes is caused by a combination of lack of memory and lack of comparison. Then the sensitivity to Gist Changes will still be better than sensitivity to Feature Changes, but sensitivity to Feature Changes will be better in the 4AFC task than in the Same/Different task.

Method

Participants. Seventy-nine students from the Radboud University Nijmegen (59 female, 20 male, average age 20.4) participated and received 5 euros for participation. The experiment was part of a computerized questionnaire containing several unrelated decision tasks. Participants were randomly assigned to either the Same/Different task (47 participants) or the 4 AFC task (32 participants).

Stimuli. There were 10 test sets of images; 8 "critical" sets and 2 filler sets¹⁷. For the Same/Different task the critical test sets consisted of a Target image, a Gist Change image (mean size of change 4% of the image) and a Feature Change image (mean size of change 17 % of the image). For the 4AFC task the critical test sets consisted of a Target, Gist Change and Feature Change images and an image in which both the Gist Change and the Feature Change were present (Combined Change image). For the filler sets the Gist Change was not perceived as a gist change by the target group, and were therefore considered invalid (see chapter 4). The results for the filler sets are not reported. For an example of a critical test set see Figure 6.1.

¹⁷ Critical test sets were test sets 1-3 and 6-10 in Figure 4.2. Test sets 16 and 18 in Figure 4.2 served as filler sets.

Design and Procedure. The experiment consisted of a study phase and a test phase. In the study phase 10 images (the Target images) were presented sequentially on a computer screen for 5 seconds each. Each image was followed by a blank screen for 1 second. The test phase consisted of 10 trials. Each trial corresponded to 1 of the 10 studied images. In the Same/Different task participants had to respond whether they thought a test image was exactly the same or different, compared to the studied image. For each studied image the Target, Feature Change or Gist Change image was presented in the test phase. The test images were varied among participants, so that each participant on average was tested with about equal frequencies of each of the three types of images.

In the 4AFC task each trial consisted of a choice between four images displayed on one screen. The alternatives always consisted of four types of image: the Target, Gist Change, Feature Change and Combined Change images. The Combined Change images were included to ensure that a decision could not be based on information available in the test phase instead of information retrieved from memory. Without the Combined Change image participants might correctly infer that the Target image was the image similar to both distractor images except for one aspect, whereas the distractor images differed from each other on two aspects. The position on screen for each of the four types of images varied between trials. Participants were asked to click the image they thought was exactly the same as the studied image. To familiarize participants with the response procedure and the types of images to expect in the experiment, a practice trial was provided directly after the study phase. The test phase started approximately 5 minutes after the study phase. The order of the images in the study and in the test phase was randomized between participants. No information was given about the types of changes and no feedback was provided on response accuracy or correct answers.

Results

For each participant the responses per image type were combined. For the Same/Different task this resulted for each participant in a proportion correct recognition for Target images, and proportion false recognition Feature Change images and Gist Change images, each based on two or three trials. When a Feature or Gist Change image is falsely recognized as the Target, this indicates that the difference is not detected. For the 4AFC task the responses for all 8 critical trials were combined. This results in a proportion correct recognition for Target images and proportion false recognition for Feature Change, Gist Change and Combined Change images for each participant. The mean proportion correct recognition (Target) and false recognition scores for the Same/Different task and the 4AFC task are displayed in Table 6.1.



Target



Feature Change



Gist Change



Combined Change
(4AFC task)

Figure 6.1. Example of stimuli used in the experiment. Within rows the gist is the same (woman under the tree), within columns the features is the same (water level). The Combined Change image was used in the 4AFC task only.

Table 6.1. Mean proportions correct recognition of Target images and false recognition of Feature Change, Gist Change and Combined Change images in the Same/Different task and the 4AFC task

	Correct Recognition		False Recognition	
	Target	Feature Change	Gist Change	Combined Change
Same/Different (n = 47)	0.83	0.68	0.30	-
4AFC (n = 32)	0.62	0.27	0.07	0.04

In the Same/Different task Feature Change images are more often falsely recognized than Gist Change images ($t(46)=5.71, p < 0.001$). In the 4AFC task also Feature Change images are more often falsely recognized than Gist Change images ($t(31)=5.30, p < 0.001$).

The results indicate that at least some of the failure to detect Feature Changes is caused by a lack of memory for feature information. As in the 4AFC task the need for comparison of all retained information was necessary to choose the correct image, it is assumed that participants will have compared all information available from memory. If the poor detection of Feature Changes was only caused by a failure to compare, than in the 4AFC task false recognition of Feature Change images should have been equal to false recognition of Gist Change images.

Comparing results

To examine if some part of the poor detection of Feature Changes in a Same/Different task is caused by a failure to compare, the results from the Same/Different task and the 4AFC task are compared. However, these results cannot be compared directly. The main difference between the forced choice task and the Same/Different task is that the proportions in the forced choice task are not independent, whereas the proportions in the Same/Different task are. For instance, in the Same/Different task participants can be very accurate for Target images, and perform very poorly for Feature Changes. Poor performance for Feature Changes in a Same/Different task means that the Feature Change image is often falsely recognized as the Target image. In the 4AFC task, correct recognition of the target images is linked to false recognition of the other images. So when correct recognition increases in a 4AFC task, false recognition decreases.

For the comparison of the results from the Same/Different task and the 4AFC task the data obtained from the Same/Different task were recoded to serve as estimates for expected proportions. The expected proportions calculated from the Same/Different task were then compared to the observed proportions in the 4AFC task. Two assumptions underlay the recoding procedure. The first assumption is that if a participant did not detect a change, he or she would always respond Same. This implies that the correct Different responses are all based on true detection of the difference, and are therefore not a result of guessing. A correct response is therefore interpreted as memory for the changed property. The second assumption is that if a participant did not detect a change, this would be a result of no memory for the changed property. This implies that for a participant with no memory for the changed property

selection of an image in a forced choice situation would be based on guessing. Together the assumptions imply that in the Same/Different task change detection performance reflects true memory. The calculation procedure is displayed graphically in Figure 6.2.

The formulas used to arrive at an expected proportion correct recognition for the Target and false recognition for the Feature and Gist Change image are presented below. The Expected values are presented in Table 6.2. In the formulas P(F) stands for "proportion of correct memory for feature information". This corresponds to proportion correct "Different" responses for Feature Change images in the Same/Different task. P(~F) stands for "proportion no memory for feature information". This corresponds to proportion incorrect "Same" responses for the Feature Change image in the Same/Different task. The proportions "no memory for feature information" match the proportion False Recognition presented in Table 6.1. The Same notation applies to the proportions for gist. For gist, P(G) stands for "memory for gist information" and P(~G) for "no memory for gist information".

Expected proportion correct recognition Target:

$$\hat{P}(\text{select Target image}) = P(F) * P(G) + \frac{P(\sim F) * P(G)}{2} + \frac{P(F) * P(\sim G)}{2} + \frac{P(\sim F) * P(\sim G)}{4}$$

Expected proportion false recognition Feature Change image:

$$\hat{P}(\text{select Feature Change image}) = \frac{P(\sim F) * P(G)}{2} + \frac{P(\sim F) * P(\sim G)}{4}$$

Expected proportion false recognition Gist Change image:

$$\hat{P}(\text{select Gist Change image}) = \frac{P(F) * P(\sim G)}{2} + \frac{P(\sim F) * P(\sim G)}{4}$$

Expected proportion false recognition Combined Change image:

$$\hat{P}(\text{select Combined Change image}) = \frac{P(\sim F) * P(\sim G)}{4}$$

Table 6.2. Expected and Observed proportions correct recognition of the Target image and false recognition of Feature and Gist Change images for comparison between performance in the Same/Different task and the 4AFC task. The observed proportions from the 4AFC are statistically compared to the expected proportions calculated from the observed properties in the Same/Different task

	Correct Recognition		False Recognition	
	Target	Feature Change	Gist Change	Combined Change
Expected from Same/Different	0.56	0.29	0.10	0.05
Observed in 4AFC	0.62	0.27	0.07	0.04
	$t(31)=1.88,$ $p = 0.07$	$t(31)=-0.65,$ $p = 0.52$	$t(31)=-1.77,$ $p = 0.09$	$t(31)=-1.16,$ $p = 0.26$

None of the observed proportions in the 4AFC task are significantly different from the expected proportions as calculated from the results of the Same/Different task.

Memory for	Expected selection based on memory for Gist and Feature information			
Memory for Gist and Memory for Feature	Target image	Feature Change image	Gist Change image	Combined Change image
Memory for Gist, no Memory for Feature	Target image	Feature Change image	Gist Change image	Combined Change image
Memory for Feature, no Memory for Gist	Target image	Feature Change image	Gist Change image	Combined Change image
No Memory for Feature, and no Memory for Gist	Target image	Feature Change image	Gist Change image	Combined Change image
Calculation of the proportion selection of each image				
Expected to <i>select</i> Target image	$0.70 * 0.68$ =0.22			
Expected to <i>guess</i> between Target and Feature Change image	$\frac{0.70 * 0.68}{2}$ =0.24	$\frac{0.70 * 0.68}{2}$ =0.24		
Expected to <i>guess</i> between Target and Gist Change image	$\frac{0.32 * 0.30}{2}$ =0.05		$\frac{0.32 * 0.30}{2}$ =0.05	
Expected to <i>guess</i> between all four images	$\frac{0.68 * 0.30}{4}$ =0.05	$\frac{0.68 * 0.30}{4}$ =0.05	$\frac{0.68 * 0.30}{4}$ =0.05	$\frac{0.68 * 0.30}{4}$ =0.05
Expected proportions for selecting each image	0.56	0.29	0.10	0.05

Figure 6.2. Expected proportions for selection of each of the four images in a 4AFC task. The proportions are calculated from the observed performance in the Same/Different task. In the top part of this figure the grey areas indicate that the corresponding image is unlikely to be selected in a 4AFC task, considering memory. For instance, when an observer has memory for both Gist and Feature information (top row), he or she is unlikely to select any other image than the Target.

Discussion

The better performance for Gist Changes compared to the performance for Feature Changes as found in both the Same/Different task and the 4AFC task is entirely consistent with the results from the Same/Different experiments described in chapter 5. This indicated that at least part of the poor performance for Feature Changes should be attributed to sparse memory for feature information. If the poor performance for Feature Changes was only caused by failure to compare, performance for Feature and Gist Change images should have been similar in the 4AFC task. If features and gist are represented in memory with equal strength, the 4AFC task would have reduced the comparison advantage for gist. As participants in the 4AFC task were more sensitive to Gist Changes, this indicates that gist is better retained in memory than feature information.

The comparison between the performance obtained from the 4AFC task and the expected values calculated from the Same/Different task revealed no difference in performance. This implies that the ability of participants to distinguish between the Target image and the changed versions of the target image is not better when participants are forced to choose. This suggests that the change detection performance measured with the Same/Different task reflects true memory for the studied Target image. If failure to compare the retained information with the perceived information during the change detection task would have effected the performance in the Same/Different task, then the performance in the 4AFC task should have been better than the calculated expected values. As there is no evidence suggesting that the poor change detection performance for Feature Changes, compared to the change detection performance for Gist Changes, is caused by a failure to compare all available information, it can be concluded that the only reason why Gist Changes are better detected than Feature Changes is the fact that gist information is remembered better than Feature information. The comparison of the results for the 4AFC task and the Same/Different task suggests that the superior detection of Gist Changes is not a result of more comparison for gist information.

In contrast to recent evidence for preserved image representation discussed in section 6.1, the reported experiment suggests that participants use all information they remembered from the image. The results of the experiment, reported in this chapter, do not suggest that people remember more than they compare. However, it should be noted that this conclusion is based on expected proportions calculated from observed performance. If people indeed use all information available to them in a change detection task as the ones described in this thesis, this finding should be replicable in other experiments using a different method to examine true memory for images. A possible other method which can be used to examine true memory is a cueing paradigm as used by Hollingworth (2003). In such a paradigm the possibly changing object is cued by for instance an arrow. This helps participants to attend to the region of interest and should facilitate comparison. However, in such an experiment participants should be asked to provide a description of the change to control for possible guessing.

Conclusion

The experiment reported in this chapter set out to assess whether gist information is more likely to be compared during a change detection task than feature information. The experiment tested whether or not the poor detection of Feature Changes as compared to the detection of Gist Changes (demonstrated in chapter 5) should be attributed to a failure to compare feature information during the change detection task, or should be attributed to a lack of memory for feature information.

Two conclusions can be drawn from the data presented in this chapter. First, it can be concluded that the gist of an image is better remembered than feature information. Second, both retained gist information and retained feature information are equally likely to be used for change detection.

When faced with the problem of the identical twin as described in the beginning of this chapter, the chance of recognizing your friend depends on the information you retained in memory. The results reported in this chapter suggest that forced comparison of the twins would not enhance the chance of identifying your friend.

7

Conclusion

In this thesis the main objective was to gain a better understanding of the concept of gist and its role in memory for images. In section 7.1 the implications of the research reported in this thesis are discussed. In section 7.2 preliminary results from an experiment examining the role of gist in a flicker paradigm are discussed. In the final section of this chapter, the research questions stated in the first chapter are re-addressed.

7.1 Discussion and implications

The definition of gist

Probably not everyone will agree that the gist as it was defined in this thesis is indeed the gist of an image. An understandable concern other researchers might have toward the concept of gist used here is that it probably does not fulfill all the characteristics which are assigned to it over the years. Three characteristics, not completely fulfilled by the concept of gist as proposed in this thesis, are briefly addressed below. The first characteristic frequently associated with gist is that the properties carrying the gist of an image are available pre-attentively (e.g., Wolfe, 1998). This implicates that the gist of an image can be extracted from an image before all (or even any) of the displayed objects are recognized (Oliva, 2005; Rensink, 2000b; see Henderson & Ferreira, 2004, for an overview). Whether or not this is the case for the concept of gist as defined in this thesis is not clear, and should be addressed in further research.

The second characteristic associated with gist that is not completely fulfilled by the concept of gist in this thesis, is that it can be extracted from low level (physical) image properties (e.g., Oliva, 2005). A series of experiments by Oliva and collaborators resulted in identification of image properties facilitating efficient categorization of pictures (Oliva & Schyns, 1997, 2000; Schyns & Oliva, 1994; Oliva & Torralba, 2001, see Oliva 2005 for an overview). The ability to extract the (perceptual) gist of an image from low level image properties is based on regularities in characteristic properties for different types of scenes. For instance, a city scene usually contains buildings. These buildings will be represented at the low level properties by vertical lines. It is unlikely that the gist as referred to in this thesis can be extracted from low level image properties alone, as the concept of gist in this thesis is defined as an interpretation of the image. Per definition, an interpretation is not extracted, but is constructed.

The third characteristic associated with gist potentially not fulfilled by the concept of gist in this thesis, is that the gist of an image can be extracted very fast (e.g., Potter, 1976). Although Experiment 1 demonstrated that Gist Changes were detected faster than other changes, this does not necessarily imply that the gist of an image as defined in this thesis is extracted very fast. The time necessary to detect the Gist Change in Experiment 1 (around 1.6 seconds) was much larger than the 120 ms usually associated with fast gist extraction (e.g., Biederman, 1981, Potter 1976). However, the fast gist extraction found by other researchers (e.g., Rousselet, Joubert, & Fabre-Thorpe, 2005, Oliva & Schyns, 1997) usually concerns images with a clear central theme typically expressible by one word or a short description (for instance, kitchen, beach). The majority of the images used in this study display complex social situations and are therefore more difficult to interpret. The gist of the images used in this thesis might be more elaborate (cf. Davenport, 2005) than the gist associated with fast extraction.

Davenport (2005) shares the suggestion that fast extraction is not an essential condition for the concept of gist. In her research she compared memory for "interesting" images with memory for "control" images from the same basic-level category. An "interesting" image depicted, for instance, a cow with two extra legs growing from its neck. The corresponding "control" image displayed a normal cow. She found that when images were presented briefly (160 or 320 milliseconds), recognition memory was similar for the normal and interesting images. When images were presented longer (750 milliseconds or 2 seconds) the recognition of interesting images was better than for normal images. She concluded that people need more time to extract the more elaborate gist, and eventually memory for the elaborate gist is better.

Even though the proposed concept of gist does not fulfill all characteristics associated with the classical view of gist, it does fulfill two important characteristics. The research in this thesis demonstrated that the gist of an image is remembered and that the gist of an image is used to detect differences between images. The definition of gist in combination with the method to determine a change in gist makes it possible to predict what changes are likely to be detected.

Human visual memory and change detection

The research reported in this thesis has potentially substantial implications for research on the amount of visual information that people retain from an image. Although by itself the reported research does not directly permit conclusions on the amount of visual detail retained in memory, the research contributes to the discussion in three ways.

First, the research described in this thesis confirms the long-standing hypothesis that changes affecting the gist of an image are detected. Although this has been generally expected, this research provides the first direct evidence.

Second, the research described in this thesis demonstrates that although large changes can go undetected (see Simons & Levin, 1997), people can be sensitive to pictorially small changes as long as the change affects the perceived gist of an image. The research in this thesis also predicts and demonstrates that pictorially somewhat larger changes are difficult to detect when the change has no influence on the interpretation of the image. This has been demonstrated for different kinds of changes, both to objects of central interest and objects of marginal interest. This suggests that memory of an object of central interest does not necessarily include detailed visual information.

Third, the method to determine whether or not a change affects the gist of an image can be applied by other researchers testing memory for specific properties. For instance, when assessing if there is a difference in sensitivity for color changes and position changes, the method to determine a gist change can be applied to verify that neither the color nor the position changes systematically affect the interpretation of the image.

The effect of gist on change detection can be demonstrated even more convincingly. The Gist Changes used in the experiments were valid for a particular target group (Dutch students). By making changes which are valid Gist Changes for one target group but are Feature Changes for another target group, it can be tested if it really is the gist which makes the difference, instead of salience of the changed aspects. For instance, adding shoes in a picture of a man sitting barefooted in a mosque would probably be a Gist Change for Muslims, but not for people unfamiliar with the Islam. Such changes should be easily detected by one group, but not by the other. This test should convince critics that the effect is a result of personal interpretation and was not caused by some other factor not controlled for in the presented research which might have rendered the changed property conspicuous.

The sensitivity for Gist Changes displayed in the experiments does not necessarily imply that the same effect will be found in daily life recognition. It is possible that in daily life people are not sensitive to changes at all. As usually things do not change unexpected, people might not feel compelled to compare what they see to what they remember. Whether or not gist is always compared should be assessed in further research. A first indication for the role of gist in daily life recognition can be found in the research reported by Simons and Levin (1998). In their experiment a person was replaced by another person during a visual disruption. They found that whether or not the replacement was noticed by participants correlated with the age of the participant. Older participants were less likely to notice the change when the replaced person was "a student" than younger participants. To account for the effect of age on likelihood of detecting the change Simons and Levin (1997) initially suggest that older people encode gist different than younger people. In another experimental condition the person who was being replaced was dressed like a construction worker (Simons & Levin, 1998). In this second condition there was no difference between older and younger participants in whether or not the change was detected. The effect was explained by the fact that for younger people the replaced "student" is part of the same social group, and this is not the case for older participants, whereas the replaced "construction worker" was neither for younger nor for older participants part of their social group. The view presented in this thesis suggests that there is no difference in the way the gist is encoded, but there is a difference in the perceived gist. When the replaced person is part of the same social group as a participant the replaced person might have been interpreted differently. For instance, for an older participant the gist of the first person and the person taking his place might be: a student asking directions. Whereas for a younger participant, the gist of the first person might have been "interesting man, a potential friend" and for the second person "resembles a man I know from lecture". As for the older participants the gist of the two persons is the same, this can explain the failed detection of the change.

Gist and Content-based image retrieval

The great importance of gist for human visual memory can be seen as bad news for computer vision. Not only are computers far from able to extract meaning from images. Also, the subjective nature of gist can result in different interpretations of an image by different people and different interpretations at different times. This makes it hard to assign key words to images which will reflect the gist for every user.

The good news is that people remember more from an image than only the high-level category (tradition definition of gist). Also, the research described in this thesis demonstrates that people are able to distinguish between visually very similar images, which makes it plausible that people are very able recognize an image surrounded by numerous distractor images. Therefore, it is possible to let a computer retrieve all images from a particular high-level category (e.g., landscapes) and let the user complete the search.

An experiment conducted in our lab in which participants had to search for a previewed image by browsing through a collection of 198 thumbnail images demonstrated that people can perform such a task highly effective. On average, more than 90 % of the target images were correctly recognized. The results of the experiment also demonstrated that participants were able to search the 198 in less than 30 seconds. Interestingly, participants were faster at recognizing a target image when all surrounding distractors were thematically related (e.g., portraits, landscapes) than when the surrounding distractor images were thematically divers.

An indication for the information participants used to correctly recognize the target images in such a short time between so many distractor images was revealed by looking at the images that were falsely recognized. When an image was falsely recognized as the target image, this could suggest what information shared by the target image and the falsely recognized image was used for recognition. For instance, in one case, when the target image was an image of a harbor, the falsely recognized images were also images displaying a harbor. This suggests that the participant did remember that he or she was looking for a harbor, but did not remember the visual details necessary to distinguish between two images both depicting a harbor. Moreover, in a number of tasks, two or more participants all incorrectly identified the same image as the target. One image was even incorrectly identified as the target image by 7 out of 25 participants. In these cases the target image and the falsely recognized image could have had a very similar gist. However, this was not tested directly. The examined false recognition suggests that there is consistency in what properties of an image people remember (i.e., gist) and compare during the search task, based on which they recognize an image.

7.2 Preliminary results from the Flicker paradigm¹⁸

The experiments reported in this thesis demonstrated that people remember the gist of an image and use this retained gist information in deciding whether or not an image is exactly the same as a previously studied image. An interesting question is to what extent the superior detection of Gist Changes compared to Feature Changes in a memory difference detection task is predictive for performance for the two types of changes in a Flicker paradigm (Rensink et al. (1997)). Unlike in the previous experiments, in a flicker paradigm change detection does

¹⁸ The author would like to thank Marsha, Koen and Rob, for carrying out the experiment.

not necessarily depend on memory for the image (see Simons and Rensink (2005) and Simons and Ambinder (2005) for a discussion).

A flicker task can be used to assess what information of an image participants attend to. As the gist of an image is the most important part of the image, it is expected that the gist of an image will be first attended during a flicker task. However, as addressed in the previous section, the gist as defined in this thesis is not necessarily perceived very quickly. In a typical flicker task, the images are presented very briefly (240ms). Therefore, it is possible that participants will not be able to interpret the images in a flicker task, and subsequently, that the Gist Changes are not perceived as Gist Changes. When in a flicker task the images are not interpreted, Gist Changes are expected to be detected slower than Feature Changes, as on average the Feature Changes affect a larger area of the image than Gist Changes. To enable participants to interpret the images, and therefore perceive a Gist Change as a change affecting the gist, in one condition participants would see the original image for 5 seconds, before the flickering started.

Method. The experiment tested change detection for Gist and Feature Changes in two conditions with 12 test sets (9 critical and 3 filler test sets). Change detection performance was tested in 12 trials. Each trial consisted of two successive short (240 ms) presentations of the Target image separated by a blank screen (80 ms), followed by two successive short (240 ms) presentations of either a Feature or a Gist Change image, separated by a blank screen. The sequence was repeated until response. Two groups of 48 participants were randomly assigned to either the preview or no-preview condition. In the preview condition each trial started with a presentation of the Target image for 5 seconds, followed by a blank screen. Each participant was tested on either a Gist Change or a Feature Change image for each test set. Participants were asked to press the “space bar” on the keyboard if they detected a change in the image. For each trial the reaction time, measured from the first short presentation of the Target image until a press on the “space bar”, was recorded. Participants were asked to respond as fast and report as accurate as possible. Every time they detected a change, they wrote down the detected difference on an answer sheet. The experiment started with two practice trials. In total the experiment took about 5 minutes.

Results. Reaction times corresponding to incorrectly reported changes were disregarded. Also, reaction times over two standard deviations over the overall mean were disregarded. As there was no significant difference between the performance in the preview and no-preview condition, the reaction times are collapsed over the conditions. In Figure 7.1 the mean reaction times for the detection of Gist and Feature Changes are displayed for the nine critical test sets.

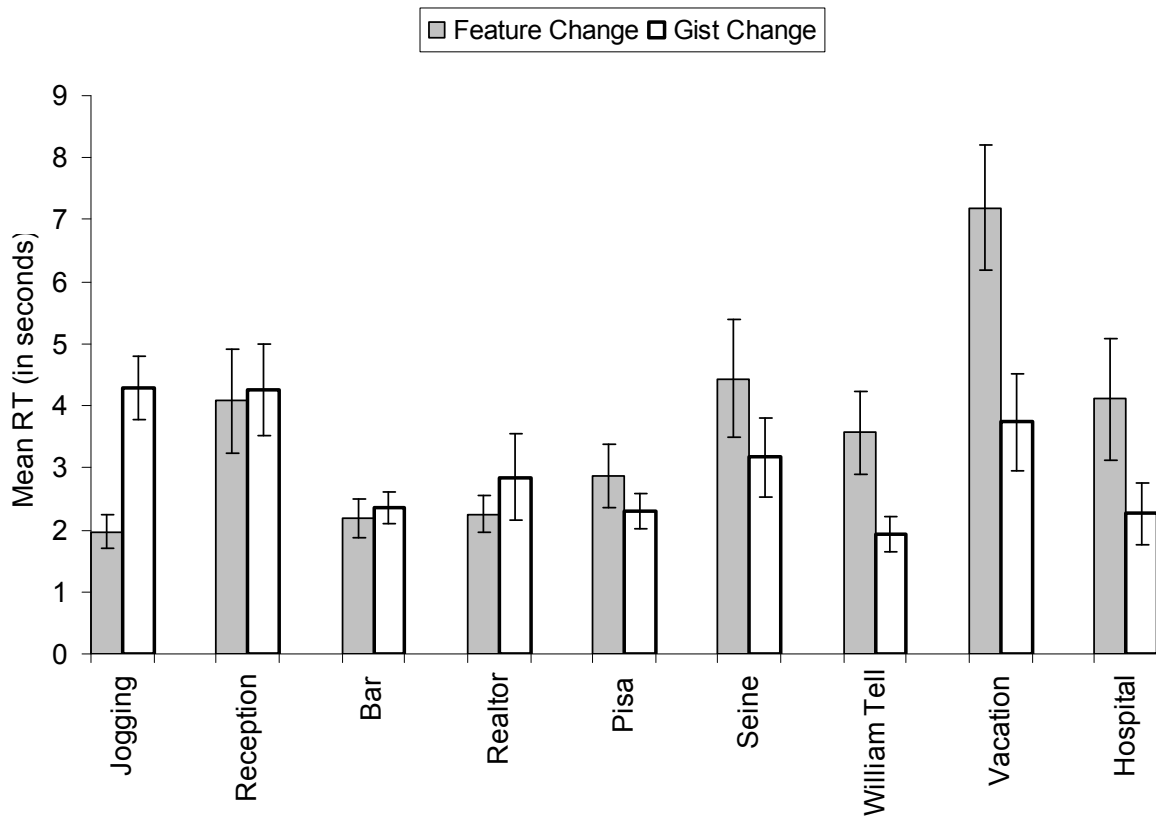


Figure 7.1. Mean reaction times (RT) for detection of Gist and Feature Changes for nine test sets in a flicker task. Error bars represent the 95% confidence interval.

Discussion. In contrast to the experiments reported in the previous chapters, in the flicker task the performance for the detection of Gist and Feature Changes does not display a clear trend. As can be seen in Figure 7.1, for one test sets participants were faster in detecting the Feature Change (test set Jogging), for five test sets participants displayed no difference in their ability to detect the two types of changes, and for three test sets participants were faster in detecting a Gist Change. The unclear pattern can not be attributed to participants inability to interpret the images as a result of the short presentation times, as in the preview condition the pattern is equally unclear. If anything, the preview of the Target image resulted in slower detection of the Feature Changes, whereas it was expected that the preview would facilitate Gist Change detection.

To test whether or not there is a facilitating effect of gist in change detection in a flicker paradigm, the experiment should be repeated with different stimuli. The stimuli used in this thesis were developed to test the role of gist in memory for images. This might have affected the visibility of the changes in a short presentation task. For example, the Gist Changes were on average smaller than the Feature Changes (see chapter 4). What can be concluded from the experiment is that difference detection performance in a memory task is not necessarily predictive of change detection in a flicker task.

7.3 Conclusion

In chapter 1 three research questions were stated to help gain a better understanding of the concept of gist and its role in memory for images. The three research questions are readdressed below.

1. *What is the gist of an image, and what are changes affecting the gist?*

The gist of an image is an interpretation of what the image is about. It reflects the essence or the meaning of the image for the observer. Changes affecting the gist of an image are changes that systematically affect the subjective interpretations of images. Whether or not a change affects the gist can be determined by comparing descriptions for the two versions of an image. When the descriptions of one version are systematically different from the descriptions of the other version, the two versions of the image are interpreted differently, and therefore have a different gist.

2. *Are changes affecting the gist better detected than other changes?*

Changes affecting the gist of an image are better (both faster and more often) detected than other changes which do not affect the gist.

3. *Is the gist of an image used more often than other image properties during a change detection task?*

As successful change detection depends on memory for the changing property and comparison of the retained information to the new situation, the better detection of Gist Changes compared to Feature Changes does not necessarily imply that memory for gist is better than memory for other image properties. Alternatively, the relatively good detection of Gist Changes might have been caused by more comparison of gist information during the change detection task. However, this was not the case in the research described in this thesis. We demonstrated that the ability of participants to distinguish between images was equally good in a change detection task as in a Four Alternative Forced Choice task. This suggests that the better detection of Gist Changes compared to Feature Changes should be attributed to better memory and not to easier comparison of gist information.

In conclusion: The subjective interpretation of an image (i.e., gist) determines what properties of an image a person remembers.

References

- Aginsky, V., & Tarr, M. J. (2000). How are different visual properties of a scene encoded in visual memory? *Visual Cognition*, 7, 147-162.
- Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, 8, 457-464.
- Angelone, B. L., Levin, D. T., & Simons, D. J. (2003). The relationship between change detection and recognition of centrally attended objects in motion pictures. *Perception*, 32, 947-962.
- Bergboer, N. H., Postma, E. O., & Herik, H. J. van den (in press). Context-based object detection in still images. *Image and Vision Computing*.
- Broek, E. L. van den (2005). Human-centered content-based image retrieval. PhD-thesis Faculty of Social Sciences, Radboud University Nijmegen, The Netherlands, Nijmegen.
- Biederman, I. (1981). On the semantics of a glance at a scene. In: M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 213-253). Hillsdale, NJ: Erlbaum Associates Inc.
- Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W., Jr. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, 103, 597-600.
- Brewer, W. F., & Treyans, J. C. (1981). The role of schemata in memory for places. *Cognitive Psychology*, 13, 207-230.
- Castelhano, M. S., & Henderson, J. M. (2005). Incidental visual memory for objects in scenes. *Visual Cognition*, 12, 1017-1040.
- Chandler, D. (1997). Visual perception 1: Searching for patterns. Downloaded from: <http://www.aber.ac.uk/media/Modules/MC10220/>
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd edition). New York; Academic Press.
- Davenport, J. L. (2005) When does a picture become interesting? *Abstracts of the Psychonomic Society*, (pp. 109). 46th Annual Meeting. Toronto, Canada.
- De Groot, A. (1965). *Thought and choice in chess*. The Hague, The Netherlands: Mouton.
- DiGirolamo, G. J., & Hintzman, D. L. (1997). First impressions are lasting impressions: A primacy effect in memory for repetitions. *Psychonomic Bulletin & Review*, 4, 121-124.
- Freeman, G. H., & Halton, J. H. (1951) Note on an exact treatment of contingency goodness-of-fit and other problems of significance. *Biometrika*, 38, 141-149.

- Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*, 316-355.
- Garner, W. R. (1974). *The processing of information and structure* (pp. 183-186). Potomac, MD: Erlbaum Associates Inc.
- Henderson, J. M., & Ferreira, F. (2005). Scene perception for psycholinguists. In J. M. Henderson and F. Ferreira (Eds.) *The interface of language, vision, and action: eye movements and the visual world*. (pp. 1-58). NY: Psychology Press.
- Henderson, J. M., & Hollingsworth, A. (2003). Eye movements and visual memory: Detection changes to saccade targets in scenes. *Perception & Psychophysics*, *65*, 58-71.
- Hollingsworth, A., & Henderson, J. M. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition*, *7*, 213-235.
- Hollingsworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 113-136.
- Hollingsworth, A. (2003). Failures of retrieval and comparison constrain change blindness in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 388-403.
- Hollingsworth, A. (2005). The relationship between online visual representation of a scene and long-term scene memory. *Journal of Experimental Psychology; Learning, Memory and Cognition*, *31*, 396-411.
- Hollink, L., Schreiber, A. Th., Wielinga, B. J., & Worring, M (2004). Classification of user image descriptions. *International Journal of Human-Computer Studies*, *61*, 601-626.
- Homa, D., & Viera, C. (1988). Long-term memory for pictures under conditions of thematically related foils. *Memory & Cognition*, *16*, 411-421.
- Intraub, H. (1980). Presentation rate and the representation of briefly glimpsed pictures in memory. *Journal of Experimental Psychology; Human Learning and Memory*, *6*, 1-12.
- Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception & Performance*, *7*, 604-610.
- Intraub, H. (1984). Conceptual masking: the effects of subsequent visual events on memory for pictures. *Journal of Experimental Psychology; Learning, Memory and Cognition*, *10*, 115-125.
- Intraub, H. (1997). The representation of visual scenes. *Trends in Cognitive Sciences*, *1*, 217-222.
- Intraub, H., & Hoffman, J. E. (1992). Reading and visual memory: Remembering scenes that were never seen. *American Journal of Psychology*, *105*, 101-114.
- Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology; Learning, Memory and Cognition*, *15*, 179-187.
- Koutstaal, W., & Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language*, *37*, 55-583.
- Levin, D. T., Simons, D. J., Angelone, B. L., & Chabris, C. F. (2002). Memory for centrally attended changing objects in an incidental real-world change detection paradigm. *British Journal of Psychology*, *93*, 289-302.

- Locher, P., Gray, S., & Nodine, C. (1996). The structural framework of pictorial balance. *Perception*, 25, 1419-1436.
- Loftus, E. F., & Ketcham, K. (1991). *Witness for the defense: The accused, the eyewitness, and the expert who puts memory on trial*. New York: St. Martin's Press.
- Loftus, G. R., & Masson, M. E. J., (1994). Using confidence intervals in within-subjects designs. *Psychonomic Bulletin & Review*, 1, 476-490.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: a user's guide* (2nd edition). Lawrence Erlbaum Associates, Inc. Mahwah, New Jersey.
- Mandler, J. M., & Johnson, N. C. (1976). Some of the thousand words a picture is worth. *Journal of Experimental Psychology; Human Learning & Memory*, 2, 529-540.
- Mandler, J. M., & Ritchey, G. H. (1977). Long-term memory for pictures. *Journal of Experimental Psychology; Human Learning and Memory*, 3, 386-396.
- McConkie, G. W., & Currie, C. B. (1996). Visual stability across saccades while viewing complex pictures. *Journal of Experimental Psychology; Human Perception & Performance*, 22, 563-581.
- Melcher, D. (2006). Accumulation and persistence of memory for natural scenes. *Journal of Vision*, 6, 8-17.
- Mitroff, S. R., Simons, D. J., & Levin, D. T. (2004). Nothing compares two views: Change blindness can occur despite preserved access to the changed information. *Perception & Psychophysics*, 66, 1268-1281.
- Mondy, S., & Coltheart, V. (2000). Detection and identification of changes in naturalistic scenes. *Visual Cognition*, 7, 281-296.
- Nelson, D. L., Reed, V. S., & Walling, J. R. (1976). Pictorial superiority effect. *Journal of Experimental Psychology; Human Learning and Memory*, 2, 523-528.
- Nickerson, R. S., & Adams, M. J., (1979). Long-term memory for a common object. *Cognitive Psychology*, 11, 287-307.
- Oliva, A. (2005). Gist of a scene. In: L. Itti, G. Rees & J. Tsotsos, (Eds.), *Neurobiology of attention* (pp. 251-256). Academic Press, Elsevier.
- 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, 72-107.
- Oliva, A., & Schyns, P. G. (2000). Diagnostic color blobs mediate scene recognition. *Cognitive Psychology*, 41, 176-210.
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42, 145-175.
- Potter, M. C. (1975). Meaning in visual search. *Science*, 187, 965-966.
- Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology; Human Learning & Memory*, 2, 509-522.
- Potter, M. C., Staub, A., & O'Conner, D. H. (2004). Pictorial and conceptual representation of glimpsed pictures. *Journal of Experimental Psychology; Human Perception & Performance*, 30, 478-489.

- Potter, M. C. (1993). Very short-term conceptual memory. *Memory & Cognition*, *21*, 156-161.
- Rensink, R. A. (2000a). Seeing, sensing, and scrutinizing. *Vision Research*, *40*, 1469-1487.
- Rensink, R. A. (2000b). The dynamic representation of scenes. *Visual Cognition*, *7*, 17-42.
- Rensink, R. A. (2002). Change Detection. *Annual Review of Psychology*, *53*, 245-277.
- Rensink, R. A., O' Regan, J. K., & Clark, J. J. (1997). To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science*, *8*, 368-373.
- Rosch, E., Mevis, C. B., Gray, W. B., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.
- Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get the "gist" of real-world natural scenes? *Visual Cognition*, *12*, 852-877.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.
- Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: evidence for time- and spatial-scale dependent scene recognition. *Psychological Science*, *5*, 195-200.
- Scott-Brown, K. C., Baker, M. R., & Orbach, H. S. (2000). Comparison blindness. *Visual Cognition*, *7*, 253-267.
- Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of verbal learning and verbal behavior*, *6*, 156-163.
- Simons, D. J. (1996). In sight, out of mind: when object representations fail. *Psychological Science*, *7*, 301-305.
- Simons, D. J., & Ambinder, M. S. (2005). Change blindness. *Current Directions in Psychological Science*, *14*, 44-48.
- Simons, D. J., & Levin, D. T. (1997). Change Blindness. *Trends in Cognitive Sciences*, *1*, 261-267.
- Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, *5*, 644-649.
- Simons, D. J., & Rensink, R. A. (2005). Change blindness: past, present, and future. *Trends in Cognitive Sciences*, *9*, 16-20.
- Simons, D. J. (2000). Current approaches to change blindness. *Visual Cognition*, *7*, 1-15
- Simons, D. J., Chabris, C. F., Schnur, T., & Levin, D. T. (2002). Evidence for preserved representations in change blindness. *Consciousness and Cognition*, *11*, 78-97.
- Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE transactions on pattern analysis and machine intelligence*, *22*, 1349-1380.
- Staalduinen, M. van, Lubbe, J. C. A. van der, & Backer, E. (2004). Circular analysis based Line Detection Filters for Watermark Extraction in X-ray Images of Etchings. *ASCI*, June 2004, 305-310.
- Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, *25*, 207-222.

- Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: single-trial learning of 2500 visual stimuli. *Psychonomic Science*, *19*, 73-74.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.
- Tatler, B. W., Gilchrist, I. D., & Rusted, J. (2003). The time course of abstract visual representation. *Perception*, *32*, 579-592.
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520-522.
- Vallacher, R. R., & Wegner, D. M. (1987). What do people think they're doing? Action Identification and Human Behavior. *Psychological Review*, *94*, 3-15.
- Wegner, D. M., & Vallacher, R. R. (1986). Action Identification. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition: Foundations of social behavior*. (pp. 550-582). New York: Guilford.
- Werner, S., & Thies, B. (2000). Is 'change blindness' attenuated by domain-specific expertise? An expert-novices comparison of change detection in football images. *Visual Cognition*, *7*, 163-174.
- Wolfe, J. M. (1998). Visual memory: what do you know about what you saw? *Current Biology*, *8*, 303-304.
- Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting memories* (pp. 71-94). Cambridge, MA: MIT Press.

Appendix

Size of affected area and Type of change

Table A.1 displays a description of the applied changes for each image and the size of the change in percentage of pixels from the whole image. The affected area of the changes was calculated by determining for each pixel whether the pixel was the same in the Original image and the changed images.

Table A.1. Types of changes and the percentage of pixels of the Original image which were different after a Gist Change and after a Feature Change for 18 test sets.

Test set	Type of change and Pixels affected by change (%)			
	Gist Change		Feature Change	
1. Couple	Addition, fangs	>1	Color, hair man and woman	21
2. Bar	Replacement, woman-man	11	Replacement, women-women	10
3. Pisa	Changed, pointing-photographing	10	Mirroring, tower	27
4. Realtor	Replacement, for sale-sold	21*	Replacement, house-house	31
5. Jogging	Addition, knife	1	Deletion, lamp	2
6. Hospital	Replacement, book-clipboard	4	Color, sheets, curtains, wall	28
7. Reception	Deletion of cast and crutch	2	Addition of water cooler	2
8. Seine	Mirroring, women	1	Replacement, riverbank-water	19
9. William Tell	Deletion, apple	>1	Addition, knight	23
10. Vacation	Replacement, deer-kangaroos	4	Deletion, clouds	9
11. New year	Changed, 5 to 12- 5 past 12	>1	Replacement, couple-couple	35
12. Doctor	Deletion, suction cup dart	>1	Replacement, doctor-doctor	20
13. Gambling	Deletion, poker chips	1	Replacement, man-women	9
14. Campfire	Addition, Loch Ness monster	5	Deletion, tree	10
15. Paris	Deletion, Eiffel tower	1	Deletion, parasol	4
16. Computer	Addition, wheelchair	2	Addition, books	2
17. Car crash	Replacement, stethoscope-hammer, pliers	1	Color, cars	8
18. Still life	Deletion, flame	>1	Deletion, banana	12

* The complete image had shifted a little. This resulted in a higher percentage of affected pixels than would be suggested by the change.

Position in image

The position in the images of the area affected by the changes was determined by imposing a 4 by 4 grid on each image. Each of the grid cells in which a part of the change was visible was counted (for images from the ten valid test sets). The percentage of changes within a cell is displayed for the Gist Change and the Feature Change.

Gist Change				Feature Change			
0	4	4	0	11	8	6	5
0	4	11	4	9	11	5	3
7	11	22	11	8	8	5	5
0	11	7	4	5	6	5	3

(a) (b)

Figure A.1 Percentage of parts of changes visible in each position in an image for Gist Changes in (a) and for Feature Changes in (b).

Summary

Although people experience a fully detailed view of the world, people are surprisingly blind to changes. On the other hand, people are capable to recognize a great numbers of images which they have seen years ago. This apparent contradiction inspired research on what people remember from images. Theoretical positions on the extent to which visual information is retained in memory vary widely. Nonetheless, there seems to be a consensus that at least the *gist* of an image is retained in memory. The research reported in this thesis provides a better understanding of what people remember from an image, the concept of gist and the role of gist in memory for images.

If the gist of an image is retained in memory, then one would expect that a change affecting the gist is detectable by people. Although researchers worked under this assumption for years, direct evidence that supports the prediction (that people detect changes affecting the gist) is lacking. To provide evidence for the facilitating effect of gist on difference detection a firm understanding of the concept of gist is necessary. A better understanding of the concept of gist and its role in memory for images is achieved in two stages. In the first stage (chapters 2-4) the definition of gist is addressed and a method is proposed to determine whether a change to an image affects the gist of the image. In the second stage (chapters 5 and 6) experiments are reported which aim to test the role of gist in difference detection based on memory for the image.

The gist of an image is commonly defined as a high level description of the essence or the central theme of the image. Over the years many characteristics have been assigned to gist. For instance, it has been assumed that the gist of an image is perceived very quickly, is retained in memory and is used for difference detection. In this thesis an additional characteristic is assigned to gist: gist is subjective (i.e., the perceived gist can vary over individuals and contexts). The concept of gist was redefined in such a way that the new definition stipulates the subjective nature of gist and reflects what people are likely to remember from an image. The definition of gist stated in this thesis is: gist is a subjective interpretation of what an image is about. Following predictions, derived from Action Identification Theory, it is expected that people will interpret an image at the highest possible level of interpretation which includes those aspects that are considered meaningful to the perceiver.

The definition of "subjective gist" provoked the need for a method to objectively determine whether or not a change to an image affects the subjective gist of the image. In chapter 3 a method is proposed which makes it possible to determine a "Gist Change". This method

determines a Gist Change in three steps. In the first step two groups of participants (Generators) provide descriptions for images reflecting their interpretation of the gist of an image. One group describes the original image, the second group describes the altered version of the image. In the second step different participants (Raters) judge the appropriateness (good fit/poor fit) of the descriptions for each of the interpreted images (the original image and the altered version of the image). In the third step the researcher analyses whether or not the two groups of descriptions provided by the Generators are significantly different, based on the appropriateness scores assigned by the Raters. When the appropriateness scores assigned for the two groups of descriptions are significantly different this indicates that the associated images are interpreted differently by the Generators. The method makes it possible to determine whether or not a change affects the gist of an image without the need to determine exactly how people interpret a particular image.

To be able to test whether the gist of an image is stored in memory and used in a difference detection task, test stimuli were developed. Each test set consists of an original image and two altered versions of the original image. For one of the altered versions the change was supposed to affect the gist of the image (Gist Change image). For the second altered version the change was supposed to not affect the gist (Feature Change image). To ensure the validity that changes in a difference detection task are a result of gist, several confounding factors are addressed. For instance, to eliminate a possible effect of the size of the area affected by a change, the gist changes were relatively small compared to the feature changes. Further, the proposed method (chapter 3) was applied to the test stimuli to ensure that the gist change was perceived as a gist change by the target group and that the feature change was perceived to have no effect on the gist. Ten test sets were determined to be suitable for examining the role of gist in difference detection.

The validated images were used in three experiments. Experiment 1 demonstrated that participants are able to detect a relatively small change to an image when the change affects the gist of the image. Additionally Experiment 1 demonstrated that Gist Changes are more easily detected than Feature Changes. Experiment 2 demonstrated that participants encode the gist of an image in memory without being specifically told that memory for the presented images will be tested. This provides evidence that the gist of an image is encoded in memory incidentally. Comparison of the results from Experiment 1 and 2 suggests that gist information was less affected by the incidental encoding task than feature information. Experiment 3 set out to test whether the superior detection of Gist Changes should be attributed to *better memory* for gist information compared to feature information, or should be attributed to *more comparison* of gist information than of feature information during the difference detection task. To this effect, recognition for studied images was tested in a Four Alternative Forced Choice (4AFC) task. It is assumed that in a 4AFC task the side by side display of all versions of an image will increase the salience of the changed areas, and subsequently forces participants to use all information available to them from memory to identify the studied image. The results from the 4AFC task were compared to the "expected results" obtained from observed performance in the Old/New task condition in Experiment 3. The analysis revealed that the superior detection of Gist Changes compared to Feature Changes is a result of better memory for gist information, and is not a result of failure to compare feature information during an Old/New difference detection task.

In conclusion, the research reported in this thesis confirms the long-standing assumption that changes affecting the gist of an image are easily detectable. The research addresses the important characteristic of gist that previously received little attention; the gist of an image is subjective. Moreover, by devising a method to determine objectively whether or not a change to an image affects the gist, the ability is created to test whether the subjective gist has an effect on difference detection. The research provides the first direct evidence that people remember the gist of an image and that the encoded gist information is used to detect differences between images.

Samenvatting

Hoewel mensen een zeer gedetailleerd beeld van de wereld waarnemen, zijn mensen verrassend blind voor veranderingen. Anderzijds zijn mensen erg goed in staat om grote aantallen plaatjes te herkennen die zij jaren geleden gezien hebben. Deze ogenschijnlijke tegenstelling inspireerde onderzoek naar wat mensen onthouden van plaatjes. Theorieën betreffende de mate waarin visuele informatie in het geheugen wordt opgenomen verschillen sterk. Er is niettemin consensus dat ten minste de "gist" (kern, hoofdgedachte, betekenis, clou, essentie) van een plaatje wordt onthouden. Het onderzoek dat in dit proefschrift wordt beschreven geeft een beter inzicht in wat mensen onthouden van een plaatje, het concept "gist" en de rol van "gist" in dit visuele geheugen.

Als de "gist" van een plaatje wordt onthouden, dan zou men verwachten dat een verandering, die de "gist" van een plaatje beïnvloedt, door mensen opgemerkt zal worden. Hoewel onderzoekers jarenlang onder deze veronderstelling werkten, ontbreekt rechtstreeks bewijs voor deze stelling (dat mensen veranderingen, die de "gist" beïnvloeden, zullen waarnemen). Om te kunnen aantonen dat mensen inderdaad veranderingen, die de "gist" van een plaatje beïnvloeden, zullen detecteren is een duidelijk inzicht in het concept "gist" noodzakelijk. Een beter inzicht in het concept "gist" en de rol van "gist" bij het onthouden van plaatjes wordt bereikt in twee stappen. In de eerste stap (hoofdstukken 2-4) wordt de definitie van "gist" behandeld en wordt een methode voorgesteld om te bepalen of een verandering in een plaatje de "gist" van het plaatje beïnvloedt. In de tweede stap (hoofdstukken 5 en 6) worden experimenten behandeld die testen wat voor rol de "gist" speelt in verschil detectie.

De meest gebruikelijke definitie van de "gist" van een plaatje is: een beschrijving op hoog/abstract niveau van de essentie of het centraal thema van een plaatje. In de loop van de jaren zijn vele kenmerken toegewezen aan "gist". Bijvoorbeeld, men veronderstelt dat de "gist" van een plaatje zeer snel wordt waargenomen, wordt onthouden en wordt gebruikt voor het detecteren van verschillen tussen plaatjes. In dit proefschrift wordt een extra kenmerk toegewezen aan "gist": "gist" is subjectief (d.w.z., de waargenomen "gist" kan variëren per persoon en/of context). Het concept "gist" wordt in dit proefschrift opnieuw gedefinieerd waarbij de nieuwe definitie de subjectieve aard van "gist" benadrukt en verwijst naar wat mensen waarschijnlijk zullen onthouden van een plaatje. De definitie die in dit proefschrift wordt gebruikt is: "gist" is een subjectieve interpretatie van de essentie van een plaatje. Op basis van de bestaande "Action Identification Theory", wordt verwacht dat mensen een plaatje op het hoogst mogelijke (abstractie)-niveau zullen interpreteren, en daarbij die aspecten betrekken die door de waarnemer als zinvol worden beschouwd.

De definitie van "subjectieve gist" maakt een methode om objectief te bepalen of een verandering in een plaatje al dan niet de "gist" van het plaatje beïnvloedt des te noodzakelijker. In hoofdstuk 3 wordt een methode voorgesteld die het mogelijk maakt om een "Gist verandering" te bepalen. Deze methode bepaalt een "Gist verandering" in drie stappen. In de eerste stap leveren twee groepen proefpersonen (de Generators) beschrijvingen voor plaatjes die hun interpretatie van de "gist" van het plaatje weergeeft (een groep beschrijft het originele plaatje de andere groep beschrijft het veranderde plaatje). In de tweede stap bepalen andere proefpersonen (de Raters) of de beschrijvingen passend zijn voor elk plaatje (originele plaatje en het veranderde plaatje). In de derde stap analyseert de onderzoeker of de door de Raters toegekende scores (beschrijving past goed of past minder goed) voor de door de Generators geleverde beschrijvingen al dan niet significant verschillend zijn voor de twee groepen beschrijvingen (beschrijvingen van het originele plaatje en beschrijvingen van het veranderde plaatje). Wanneer de toegekende scores voor de groepen beschrijvingen beduidend verschillend zijn wijst dit erop dat de bijbehorende plaatjes verschillend worden geïnterpreteerd door de Generators. De methode maakt het mogelijk om te bepalen of een verandering in een plaatje de "gist" van het plaatje beïnvloedt, zonder dat het nodig is om precies te bepalen hoe mensen dat plaatje interpreteren.

Om te kunnen testen of de "gist" van een plaatje wordt onthouden en wordt gebruikt voor verschil detectie werden teststimuli ontwikkeld. Elke set van testplaatjes bestaat uit een origineel plaatje en twee veranderde versies van het originele plaatje. Bij één van de veranderde versies wordt verondersteld dat de "gist" van het plaatje is beïnvloed ("gist" Verandering). Voor de andere veranderde versie wordt verondersteld dat de "gist" juist niet is beïnvloed, dus uitsluitend de kenmerken van het plaatje (Feature Verandering). Om ervoor te zorgen dat een verschil in detectie van de veranderingen van "Gist Verandering" in vergelijking met "Feature Verandering" in een verschil detectie opdracht uitsluitend te wijten is aan de "gist", zijn verscheidene mogelijk interfererende factoren besproken. Om ongewilde nadruk te elimineren is, bijvoorbeeld, de grootte van het veranderde gebied bij "Gist Veranderingen" relatief klein in vergelijking met de "Feature Veranderingen". Verder, werd de voorgestelde methode (hoofdstuk 3) toegepast op de teststimuli om te garanderen dat de "Gist Verandering" als "gist" verandering wordt ervaren door de doelgroep en dat voor de Feature Verandering geen verandering in "gist" werd ervaren. Tien setjes testplaatjes werden geschikt bevonden om de rol van "gist" in verschil detectie te onderzoeken.

De gevalideerde plaatjes werden gebruikt in drie experimenten. Experiment 1 toonde aan dat proefpersonen een relatief kleine verandering in een plaatje kunnen ontdekken wanneer de verandering de "gist" van het plaatje beïnvloedt. Verder toonde Experiment 1 aan dat "Gist Veranderingen" gemakkelijker werden gedetecteerd dan "Feature Veranderingen". Experiment 2 toonde aan dat proefpersonen de "gist" van een plaatje onthouden zonder dat hen specifiek wordt verteld dat hun visuele geheugen zal worden getest. Dit levert bewijs dat de "gist" van een plaatje terloops wordt opgeslagen/gecodeerd. Vergelijking van de resultaten van Experiment 1 en 2 suggereert dat "Gist"-informatie bij de terloopse geheugenopslag beter onthouden worden dan de "Feature"-informatie. In Experiment 3 werd getest of de betere detectie van "Gist Veranderingen" verklaard kan worden vanuit het feit dat mensen de "Gist" beter onthouden dan de "Features", of dat mensen tijdens een vergelijkings opdracht de onthouden "gist" laten prevaleren boven andere onthouden kenmerken van een plaatje. Om dit te onderzoeken werd herkenning van bestudeerde plaatjes getest in een Vier Alternatieven

Gedwongen Keuze taak (4AFC). Er wordt verondersteld dat in een 4AFC taak het naast elkaar tonen van de verschillende versies van een plaatje de veranderingen in het oog springen en proefpersonen daardoor dwingt om alle onthouden informatie te gebruiken om het bestudeerde plaatje te identificeren. De resultaten van de 4AFC taak werden vergeleken met "verwachte resultaten" die berekend werden uit waargenomen prestaties in de Oud/Nieuw taak conditie in Experiment 3. De analyse toonde aan dat de betere detectie van "Gist Veranderingen" in vergelijking met "Feature Veranderingen" een resultaat is van beter geheugen voor "gist"-informatie en niet het onvoldoende benutten van opgeslagen "Feature"-informatie bij de Oud/Nieuw verschil detectie opdracht.

Samenvattend, bevestigt het onderzoek, beschreven in dit proefschrift, de al lang bestaande veronderstelling dat veranderingen die de "gist" van een plaatje beïnvloeden, gemakkelijk detecteerbaar zijn. Verder behandelt het onderzoek een belangrijk kenmerk van "gist" dat eerder weinig aandacht kreeg, namelijk dat de "gist" van een plaatje subjectief is. Door een methode te ontwikkelen om objectief te bepalen of een verandering in een plaatje al dan niet de "gist" beïnvloedt, was het mogelijk te testen of "subjectieve gist" een rol speelt bij verschil detectie. Het onderzoek levert het eerste directe bewijs dat mensen de "gist" van een plaatje onthouden en dat de onthouden "gist" wordt gebruikt om verschillen tussen plaatjes te ontdekken.

Biography

Xandra van Montfort was born on the 7th of March, 1974, in Peize (Drente). From 1986 to 1994 she attended the "Bernardinus College" in Heerlen, where she obtained her HAVO and VWO diplomas in respectively 1992 and 1994.

In 1994 she started her study Technology Management with the technical component Building Engineering at the Technische Universiteit Eindhoven. During her study she was introduced to the fascinating world of experimental research. In a student project on automatic human behavior she found that people speak more softly when primed with a picture of a library than when primed with a picture of a funfair. The seed for scientific research was planted. During her graduation project under supervision of prof. Gideon Keren she researched the effects of memory on choice behavior. In 2000 she received a Master degree.

After researching the effects of personal values on "housing preferences" as a junior researcher at the Technische Universiteit Delft in 2001, she went back to her Alma Mater in Eindhoven to start as a PhD student in 2002. The result is presented here.