GEO 511 Master's Thesis

An Empirical Assessment of Correction Methods for the Terrain Reversal Effect in Satellite Images

GEO 511 Master's Thesis

Author: Gianna Hartung
09-734-286

Supervisor:
Dr. Arzu Çöltekin

Faculty representative:
Prof. Dr. Sara Irina Fabrikant

Submission: 09/30/2016

Geographic Information Visualization & Analysis (GIVA)
Department of Geography
University of Zurich

Abstract

The use of satellite images is for many people a common tool to receive information about their environment. The correct interpretation of land forms and land cover classes is thereby crucial for an unrestricted use of satellite images. In these images, sometimes a phenomenon called "terrain reversal effect" occurs. This perceptual phenomenon affects individuals as it causes an inverted perception of three-dimensional forms –e.g. a valley can appear as a ridge and vice versa. The incidence of this effect is related to the direction of illumination in satellite maps and to several priors that influence human's interpretation of depth. Previous research suggests several methods to correct the terrain reversal effect. Many of these methods rely on the combination of a satellite image with a shaded relief map (SRM) in order to change the lighting direction. Other methods have integrated strong depth cues such as stereopsis or motion in depth perception and terrain correction methods. Although often theoretically discussed, only a few user studies in relation to correction methods for the terrain reversal effect in satellite images have been conducted until this day. The aim of this thesis is to empirically evaluate different correction methods by testing them with various land form and land cover tasks. The used correction methods are based on a SRM-overlay and several combinations including labelling, stereopsis and motion. These visualizations were tested for accuracy, response time, confidence, preference and quality in a controlled lab study. The results revealed that all corrections significantly improve the ability to detect land forms correctly, but also significantly reduce the perception of the land cover. Confidence and response time do not differ significantly between the use of the original satellite image and the correction-adapted visualizations. Furthermore, the participants' preference stands in contrast to their performance with all visualizations. Based on the results, recommendation for further investigations concerning the correction of the terrain reversal effect can be drawn.

Keywords

terrain reversal effect, satellite images, correction methods, shaded relief map, stereoscopy, motion, handedness, accuracy, response time, confidence, preference, quality

Acknowledgement

First, I would like to thank my supervisor Dr. Arzu Çöltekin for her support through the whole preparation of this thesis. I appreciated the helpful talks and constructive feedbacks a lot.

I am grateful to all the participants who took part in the experiments and therefore made this thesis possible.

Furthermore, I am thankful to Dr. Philip Jörg from the NPOC for giving me helpful hints for the design of the visualizations. I also thank Dr. Robert Hess and Dr. Jeremy Cooperstock from the McGill University of Montreal who kindly let me use their stereo test. I also thank Julien Biland for providing me some of his stimuli for the second experiment.

I would like to give special thanks to Vanessa Metz and Marcel Deggeller for reading and correcting the thesis. I also thank my sister Cristina for participating in the pilot study and helping me recruit participants.

And finally, I am thankful to my parents for making my university studies possible and supporting me for all these years.

Table of Contents

List of Figures

List of Tables

Abbreviations

2D	two-dimensional
3D	three-dimensional
a.s.l.	above sea level
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
DEM	digital elevation model
df	degrees of freedom (statistics)
EGM	earth gravitational model
F	F-value (statistics)
GDEM	Global Digital Elevation Model
IHS	intensity-hue-saturation
IQR	interquartile range (statistics)
KDE	kinetic depth effect
LC	land cover
M	mean (statistics)
Mdn	median (statistics)
MWF model	modified weak fusion model
MLE	maximum likelihood estimation
n / N	sample size (statistics)
OLI	Operational land Imager
NPOC	Swiss National Point of Contact for Satellite Images
p	p-value (statistics)
PCA	principal component analysis
r	effect size (statistics)
RGB	red-green-blue
SD	standard deviation (statistics)
SE	standard error (statistics)
SIDQT	shift invariant wavelet transform method
SPOT	Satellite pour l'observation de la terre
SRM	shaded relief map
SRTM	Shuttle Radar Topography Mission
t	t-value (statistics)
TIRS	Thermal Infrared Sensor
TM	thematic mapper
TRE	terrain reversal effect
UTM	universal transverse mercator
WGS	world geodetic system

1 Introduction

Nowadays, the technological progress and the Internet offer new possibilities for exploring the environment using maps and other visualizations to a large range of individuals. Among the widely accessible geographic visualizations, we also find satellite images and shaded relief maps (SRM), which are often used for many tasks and decisions (Çöltekin, Lokka, & Boér, 2015). Satellite images and shaded relief maps are used, for example, to explore the environment (e.g. with Google Maps), to observe landscape change and to extract paths or for monitoring purposes. These are only some of the tasks that require satellite images (e.g. Taylor, Brewer, & Bird, 2010; Toppe, 1987). Shaded relief maps are also used in a large number of domains such as visual extraction of topography, hydrology or glaciology (e.g. Jenson, 1991; Miller et al., 2012). With the spread of publicly available map providers, satellite images and relief models are used by many people with different amount of expertise in map and image interpretation. For the correct identification of spatial relationships in geographic visualizations, it is crucial that all individuals can interpret shaded relief maps and satellite images correctly (Saraf, Sinha, Ghosh, & Choudhury, 2007).

In this thesis, the focus will be on satellite image interpretation. Satellite images are two-dimensional images, but they depict three-dimensional objects. The human visual system has a number of mechanisms to perceive three-dimensional information (depth) in everyday life based on both monocular and stereoscopic cues. However, the task of perceiving three dimensions in satellite images is mainly based on shadows, and shadows can introduce complications. One difficulty that is related to perceiving depth in satellite images (and in shaded relief maps) based on shadows is the so-called "terrain reversal effect". The terrain reversal effect is a perceptual phenomenon where a three-dimensional (3D) shape is inversed, i.e., convex shapes are perceived as concave and vice versa. This effect occurs in various geographic visualizations like shaded relief maps, contour maps and satellite images (Bernabé-Poveda & Çöltekin, 2014). It is assumed that the phenomenon is produced by the direction of lighting which does not correspond to the expected light direction of the human brain.

Different terms exist in literature describing the same effect: relief inversion (e.g. Bernabé-Poveda, Callejo, & Ballari, 2005), terrain reversal effect (e.g. Bernabé-Poveda, Sánchez-Ortega, & Çöltekin, 2011), relief inversion fallacy (Kettunen & Sarjakoski, 2009) or false topographic perception phenomenon (FTTP; Saraf, Das, Agarwal, & Sundaram, 1996). When the effect is observed in visualizations where the position of the light source can be controlled (e.g., shaded relief maps), the direction of the illumination can be adapted artificially. However, in satellite images, this is not possible. Most satellites have (quasi-)polar orbits and pass the earth at a certain time (e.g. Saraf et al., 1996). Therefore, all the satellite images are illuminated from a fixed position and the terrain reversal effect can occur. The terrain reversal effect as a perceptual phenomenon is not fully explored up to now. For example, it is not completely explained to which extent the visual perception of depth is a biological trait or experience-related (Adams, Graf, & Ernst, 2004; Sun & Perona, 1998). Additionally, it is subject of discussion, if and how individual and cultural factors such as handedness or writing direction have an influence on our visual perception.

The terrain reversal effect is prevalent in satellite images and occurs everywhere on the globe, but much more commonly on the northern hemisphere. The effect is related to the actual illumination direction which opposes the direction that our visual system would expect. Human beings expect the light to come from above and whenever this is not the case, the depth perception becomes ambiguous or reversed (Gerardin, Kourtzi, & Mamassian, 2010). Different methods for correcting the terrain reversal effect in satellite images have been proposed, and many of them rely on the adaptation of the illumination direction. However, most of them introduce new perceptual challenges.

Knowing how humans perceive depth is essential for understanding the terrain reversal effect and the proposed correction methods. Depth cues are information elements (such as perspective, relative size, shading/illumination, stereopsis), which help to perceive depth, also in two-dimensional (2D) images (Goldstein, 2002). Both physiological and psychological cues help humans perceive depth in 2D images. Depth cues are integrated in a way that the information content about depth in an image can be maximized. Different depth cues have different strengths, and are differently weighted depending on the observer and the image characteristics (Bülthoff & Mallot, 1988; Landy, Maloney, Johnston, & Young, 1995; Lovell, Bloj, & Harris, 2012). Depth cues, therefore, are essential to understanding three-dimensional perception in satellite images, and can be integrated in correction methods for the terrain reversal effect.

The need to correct this effect in geographic visualizations, such as satellite images and shaded relief maps, is of interest to both providers and users. While correcting for this effect seems essential for the visual inspection and interpretation of satellite imagery, many of the suggested correction methods have not yet been tested in user studies. Whether they really help or introduce new problems has to be investigated. This project will address this gap through a comprehensive empirical evaluation of various correction methods that are proposed in literature. Most correction methods found in literature are developed based on theoretical considerations (e.g. Bernabé-Poveda, Callejo, & Ballari, 2005; Bernabé-Poveda, Callejo, & Ballari, 2005; Saraf, Das, Agarwal, & Sundaram, 1996; Saraf, Ghosh, Sarma, & Choudhury, 2005; Saraf et al., 2007; Wu, Li, & Gao, 2013). More specific aims of the project are to gain more knowledge about the illusion, and empirically assess a selected set of promising correction methods to understand which one fixes the terrain reversal effect in a way that helps with selected image interpretation tasks, and fits best for satellite map users with different expertise.

To achieve the goals mentioned above, two experiments were conducted to evaluate the selected correction methods in this project (**Figure 1.1**). The first experiment was designed as an online study that assessed a specific correction method which seemed particularly promising based on the literature review. This method overlays a satellite image with a semi-transparent shaded relief map (SRM-overlay method, see section 2.5 for further explanation) at different opacity levels (Bernabé-Poveda & Çöltekin, 2014; Gil, Arza, Ortiz, & Avila, 2014). The findings in the first experiment serve as a basis for the second one, where more correction methods are included for control and comparison. In the first experiment, the alternative visualization types (SRM-overlays with varying opacity levels) are presented to the participants and are tested on two tasks. These two tasks represent basic image interpretation tasks and include the recognition of both land forms and land cover. The second experiment was conducted as a controlled laboratory study. Its goal was to assess how the most promising variant of the SRM-overlay correction method from the first experiment compares to seven other alternatives. This includes some combinations that are uniquely proposed in this thesis. Specifically, participants' accuracy, response time and confidence were assessed as they worked with the provided correction methods (visualization types). Also, the preference and quality ratings of the participants per visualization type were evaluated. The 'best' correction method would result in a high performance (high accuracy and response time) and a positive subjective experience (confidence, preference, quality). If the participants perform equally well with the different visualizations, it might be critical to analyze subjective factors in order to give a recommendation

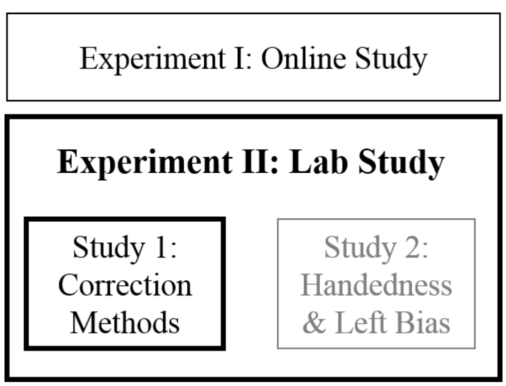


Figure 1.1: Overview of the complete experimental design. Study 1 of Experiment II is the center of interest in this thesis.

about the suitability of the correction method (Brügger, Fabrikant, & Çöltekin, 2016; Hegarty, Smallman, Stull, & Canham, 2009; Smallman, Cook, Manes, & Cowen, 2007).

The thesis follows the conventional structure of a research paper. The second chapter gives an overview over the current state of research and aims to set theoretical fundaments for the research questions and the following experiments. In Chapter 3 the research questions and hypotheses are formulated. The methodology, results and discussion are separated into two parts according to the two experiments. The three following chapters describe the material and methods used for Experiment I (Chapter 4), the results from this experiment (Chapter 5) and a short discussion of Experiment I (Chapter 6). The subsequent chapters cover Experiment II: material and methods are described in Chapter 7, the results are described in Chapter 8 and discussed in Chapter 9 including the limitations of the second experiment. Finally, the outcomes are summarized and aspects of further research are shown in Chapter 10. In the Appendix, an extended version of the material and methods can be found.

2 Literature Review

The following chapter gives an overview of fundamental and current research in the fields that were introduced in the first chapter. First, the role of satellite imagery as a source of geographic information is shortly discussed. In the following section, a review of the literature on the terrain reversal effect and some influencing visual assumptions are presented. Subsequently, the theoretical background for the terrain reversal effect as well as for its correction are approached by discussing human depth perception and different depth cues. Then, an overview of different methods to correct the terrain reversal effect is provided. Furthermore, handedness as a possible source of influences on the terrain reversal effect is discussed.

2.1 Satellite Imagery

Satellite images have turned from expert material to open-source products with high quality that is nowadays easily available through the Internet. Not only do they support geography-related sciences as for example earth science, resource management, urban planning or disaster rescue, but also do they serve non-expert users to make everyday decisions (Saraf et al., 2007).

Satellite images are acquired by different kinds of satellites. Many earth observation satellites which provide satellite imagery can circle in low earth orbits (LEO) in order to picture the earth from relative near distance. Low earth orbit satellites either fly sun-synchronous or not (Pirscher, Foelsche, Lackner, & Kirchengast, 2007). Most satellite images for scientific purposes or weather observation are captured by optical satellites that travel in a sun-synchronous orbit (e.g. Bernabé-Poveda, Sánchez-Ortega, & Çöltekin, 2011b; Bernabé-Poveda et al., 2005; Bernabé-Poveda & Çöltekin, 2014; Gil, Ortiz, Rego, & Gelpi, 2010). The sun-synchrony assures that the incidence angle of sunlight stays consistent to allow comparisons and change detection over time. The sun-synchronous satellites pass each geographical latitude two times a day (Pirscher et al., 2007). They pass the same location always at the same time in the morning and evening hours. The time when the satellites cross the equator usually is between 09:30 and 10:30 a.m. according to Saraf et al. (2005), respectively between 9:00 and 11:00 a.m. according to Gil et al. (2014). At this time, the conditions for image capturing are optimal due to little haze and good lighting conditions (Gil et al., 2014). An example of a sun-synchronous satellite is the Landsat 8. It circles on an altitude of 705 km and covers the entire globe every 16 days. The Equatorial crossing time is 10 a.m. ± 15 minutes[1].

As the satellites move, they scan the earth either from bottom up or from top down (Pirscher et al., 2007). Optical satellites usually take images in descending mode (Saraf et al., 1996). The sensor in the satellite is aligned in a way that maximal reflection from earth is perceived. The relative position of the sun to the sensor results in an optimal illumination of the satellite image. Due to the tilt of earth's rotation axis, the light direction varies between northern and southern hemisphere. Satellite images picturing the northern hemisphere are lit from south, while on images from the southern hemisphere, the light comes from north (**Figure 2.1**). The resulting 2D satellite images are orientated in a way that north is on top. For images of the northern hemisphere, this results in a lighting condition leading to misinterpretation of land forms (see section 1.2 and 1.4 for further explanations) (Saraf et al., 1996).

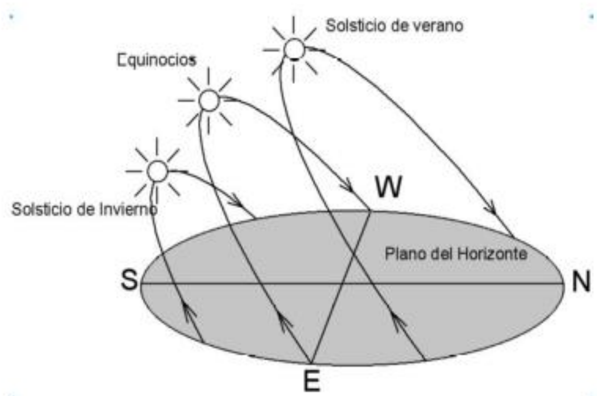


Figure 2.1: Yearly location of the sun in relation to the horizon in the northern hemisphere (Bernabé-Poveda et al., 2005).

After satellite images are collected, the further processing steps are reduced. For example, only the viewing angle of the observer can be changed afterwards, but the illumination direction in the image is fix (Saraf et al., 2005). This stays in contrast to the shaded relief models, which can be artificially lit from different directions. In relief maps as well as in satellite images, additional information about the terrain structure can be drawn from linear features such as roads or rivers (Phillips, Lucia, & Skelton, 1975). Phillips et al. (1975) observed that shaded maps improved the performance of participants when visualizing the landscape, but reading a hill shaded map might require some experience (Phillips et al., 1975).

[1] http://landsat.usgs.gov/about_ldcm.php

2.2 Terrain Reversal Effect

An essential requirement for the quality of satellite images is the ability to interpret them correctly. However, in some images this ability is disturbed by a perceptual issue called “terrain reversal effect”. This effect results in an inversion of 3D shapes, i.e. convex shapes are perceived as concave an vice versa. In

Figure 2.2: Two examples of satellite images containing the terrain reversal effect (Bernabé-Poveda & Çöltekin, 2014).

terms of satellite images for example, a valley is perceived as a ridge (Bernabé-Poveda & Çöltekin, 2014). These topographic inversions are especially harmful as most people are not aware of the existence of this phenomenon (Saraf et al., 2007) (**Figure 2.2**).

The terrain reversal effect is known for a long time, but is not completely described and explored (Bernabé-Poveda & Çöltekin, 2014). Different names have been developed: the effect is called relief inversion effect (Bernabé-Poveda et al., 2005;), terrain reversal effect (Bernabé-Poveda et al., 2011; Zhou, Zhang, & Gao, 2006), relief inversion fallacy (Kettunen & Sarjakoski, 2009), false topographic perception phenomenon (Saraf et al., 1996; Saraf et al., 2007) or pseudoscopic effect (Gil et al., 2010). In the following, the effect is called “terrain reversal effect” as this term relates to the geographic context.

In the field of geography, the terrain reversal effect is observable in artificially lit relief maps, contour maps, but also in naturally lit imagery as satellite images and aerial photographs (Bernabé-Poveda et al., 2014). Whereas in artificially lit images, human beings have the power to influence the effect, in naturally lit images, there is no direct influence possible (Bernabé-Poveda & Çöltekin, 2014). As the use of naturally lit images like satellite images is widely distributed, it may lead to an erroneous interpretation of an image or photo. Spatial relationships between land forms and image classification can also be complicated (Bernabé-Poveda et al., 2011).

The importance of the illumination direction is already discussed by Imhof (1967). Bernabé-Poveda and Çöltekin (2014) confirm that the viewing angle and the shading are critical factors for terrain perception.

This is not only true for terrain, but also for geometrical forms as discussed by Kleffner and Ramachandran (1992) as well as Liu and Todd (2004). Several researchers pointed out that an image or relief that is north-oriented, but illuminated from the south-east produces a reversal effect (e.g. Bernabé-Poveda & Çöltekin, 2014; Imhof, 1967).

Causes for the terrain reversal effect can be found in the location of shadows respectively of the lighting source in respect to the observer (Saraf et al. 2007). Other causes are different pictorial factors such as the topographic relief, the sun elevation and azimuth as well as the viewing angle, the texture of the object's slopes and the observer's position (Saraf et al., 1996, 2005, 2007; Zhang et al., 2016). For the shaded relief models, reasons for a reversed terrain can also be found in the illumination direction and the shading (Bernabé-Poveda et al., 2011, 2005; Bernabé-Poveda & Çöltekin, 2014; Gil et al., 2014; Saraf et al., 1996, 2005; Saraf et al., 2007). Although these causes are known, it is not completely understood if the terrain reversal exists because of the position of satellites and the sun or if the main cause lies in the complex visual system of human beings (Saraf et al., 2005, 2007).

The time of passage of a satellite has a large influence on the illumination of satellite images. In the morning hours, the sunlight illuminates the objects on the earth from the southern direction on the northern hemisphere whereas for the southern hemisphere the sunlight comes from north (Bernabé-Poveda et al., 2011; Bernabé-Poveda et al., 2005; Gil et al., 2014; Gil et al., 2010; Zhang et al., 2016). Therefore, three-dimensional objects on the earth's surface are represented differently in the two hemispheres as they are lit from different directions (**Figure 2.3**).

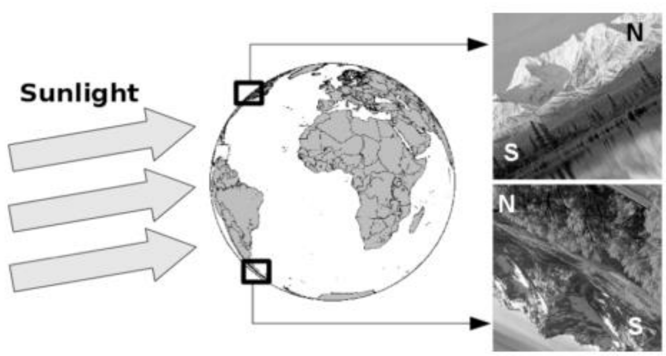


Figure 2.3: The relative position of the sun and the earth creates a different illumination in the southern and northern hemispheres (Bernabé-Poveda et al. (2011).

The terrain reversal effect occurs mostly on the northern hemisphere because of the illumination angle due to the tilted position of the earth (Bernabé-Poveda & Çöltekin, 2014; Bernabé-Poveda et al., 2011;

Bernabé-Poveda et al., 2005; Saraf et al., 2007). Gil et al. (2014) as well as Wu et al. (2013) specify that at latitudes higher than +23.5° north, the effect occurs in all images regardless of the capturing time. Also in high latitudes, the solar elevation angles are smaller and therefore, the shadows are larger on the surface (Sharma, Saraf, Das, & Baral, 2015; Wu et al., 2013). Although the phenomenon is expected to be more prevalent in mountainous terrains (Saraf et al., 1996, 2005, 2007), it also occurs in flat terrains (Meyer, 2015). Gil et al. (2014) discuss that the more rugged a terrain is, the more noticeable the effect is. It is observed that an inversion is more likely when the source of illumination is orthogonal to the linear land forms and when such land forms are surrounded by flat areas.

The terrain reversal effect is a wide spread phenomenon and not only occurs on earth, but also on Mars and on the moon (Saraf, Zia, Das, Sharma, & Rawat, 2011; Sharma et al., 2015; Wu et al., 2013). On Mars and on the moon, the effect is even more prevalent due to the lack of atmosphere, vegetation and artificial objects as well as because of the high ruggedness (Wu et al., 2013).

2.3 Depth Perception

The fact that humans perceive the terrain reversal effect relies on the mechanics of human vision and on the factors that help us perceive depth (Gil et al., 2014; Gil et al., 2010). The ability of human beings to perceive depth is an exceptional performance of our visual system and our brain. When looking for example at satellite images, most observers unconsciously evaluate the distance of various objects and try to perceive depth (Saraf et al., 1996).

Depth Cues

In order to process the 2D images that we receive from the world through our eyes into 3D images, the brain uses different hints about depth or distance called cues (e.g. Saraf et al., 2005). Depth cues can be classified in several ways. There is a main distinction between physiological (or physical) and psychological cues. This groups are called primary cues (corresponding to monocular cues) and secondary cues (corresponding to binocular cues) according to Hubona, Wheeler, Shirah, & Brandt (1999). Especially, the psychological cues play a major role in the terrain reversal effect as said by Saraf et al. (2007). The physiological cues in satellite images that influence depth perception are accommodation, convergence, binocular disparity and motion parallax. The psychological cues for the depth perception are retinal image size, linear perspective, area perspective, overlapping, shades-shadows and texture gradients (Saraf et al., 2005).

Different cues are used depending on whether the vision is binocular or monocular (e.g. Bharathi & Priyadharshini, 2016; Gil et al., 2010). Binocular cues assume that both eyes see simultaneously and that the perception of two slightly disparate images allows to perceive relief and depth. The binocular vision includes many cues that can be classified into the group of physiological cues (Gil et al., 2010). Monocular vision uses only one eye and extracts depth and relief information from psychological cues (Bharathi & Priyadharshini, 2016; Gil et al., 2010). **Table 2.1** provides an overview of different depth cues.

Table 2.1: Depth cues classified into physiological or psychological depth cue according to Mehrabi et al. 2013 and Okoshi 1976. Monocular cues are shown with grey background color.

Physiological Depth Cues	Psychological Depth Cues
Accommodation: change of the focal length of the eyes with distance	*Retinal Image Size*: previous knowledge about object's size
Convergence: angle of convergence of eyes while focusing on an object	*Linear Perspective*: the nearer to the horizon, the closer to the observer
Binocular Parallax: slightly disparate images fused into one	*Texture Gradient*: close objects appear clearer than distant ones
Depth form Defocus: different amount of blurring	*Occlusion*: Overlapping of objects
Monocular Movement (Motion) Parallax: speed changes with distance; kinetic depth perception	*Aerial Perspective*: distant objects appear hazy
	Shadowing: cast shadows of one object onto another object *Shading*: bright sides of objects are expected to be oriented towards the light source

The different depth cues have various meanings and functions depending on the context in which they are used and on the observer who uses them. In satellite images and also in shaded relief maps, shading plays an important role for depth perception. The shading cue is inseparably related to the question of the light direction; another crucial factor in satellite images and relief maps. But shading as a depth cue relies not only on the position of the light source; the position of the observer towards the object, the shape of the object and its material properties are just as important (Horn, 1989; Lovell et al., 2012; Pentland, 1989). Shading as well as texture cues are more ambiguous. They often rely on additional information as for example outlines, texture, occlusion or prior assumptions about the light source perception (Baoxia Liu & Todd, 2004; Lovell et al., 2012; Sun & Perona, 1998). An important distinction must be made between shading and (cast) shadows for depth cues. Both can influence depth perception, but mainly shading is discussed when it comes to the terrain reversal effect and its correction.

All cues have advantages and disadvantages for the perception of depth. For example, binocular disparity is a strong cue, but its quality relies on the viewing distance and the shape of the object (Lovell et al., 2012; Mehrabi, Peek, Wuensche, & Lutteroth, 2013). The characteristics of depth cues are extensively discussed by a range of literature (Goldstein, 2002; Mehrabi et al., 2013; Okoshi, 1976 Young, Landy, & Maloney, 1993).

Depth Cue Integration

Theory about depth cues implies that the combination of cues results in an enhanced knowledge about shape, distance, illumination direction etc. and therefore leads to a more accurate interpretation of images (Lovell et al., 2012). Often, more than one depth cue is available in an image and in everyday life, humans are not always aware of the prevalence of different cues (**Figure 2.4**). In the case of multiple cues, the observer of the image fuses the them in a way that maximal depth information can be extracted from the image. The question how humans integrate depth cues is not always apparent and different approaches exist to explain the human cue combination. Some of them are discussed in the following section.

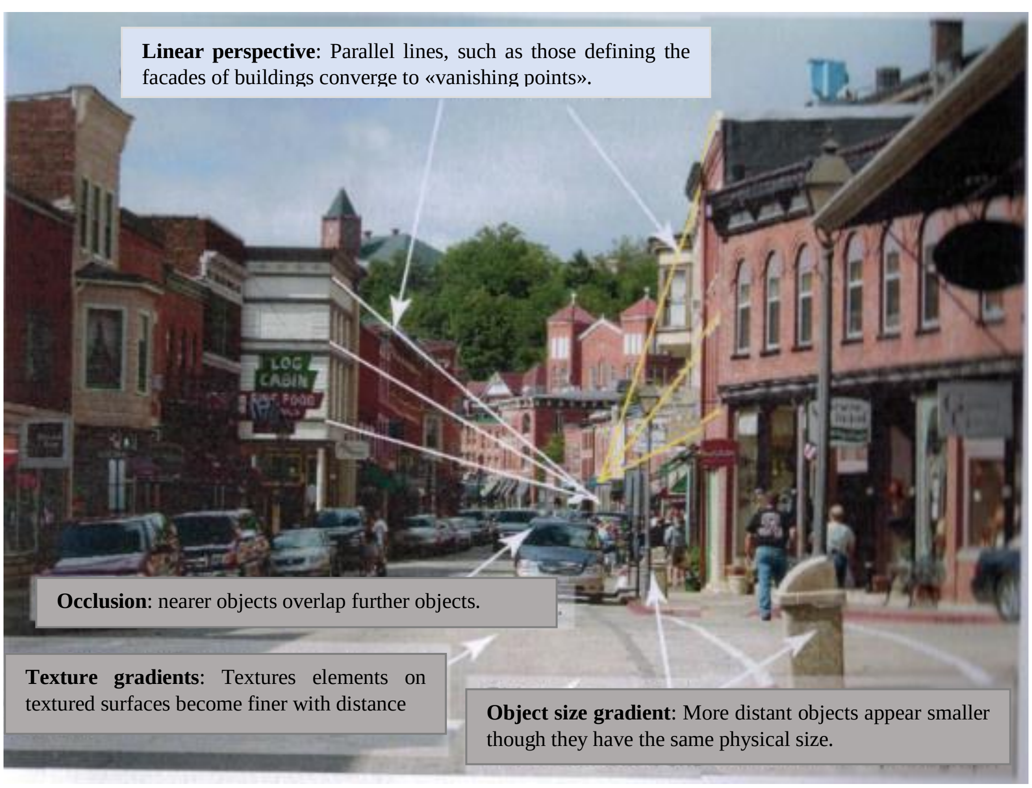


Figure 2.4: Exemplary scenery with multiple depth cues (Ware 2008).

One approach that is the prevalent assumption in current research to how cue integration works is discussed by Landy et al. (1995) with the modified weak fusion (MWF) model. This is an adaptation of the weak fusion model proposed e.g. by Maloney and Landy (1989). They suggest that as long as depth cues occur as single cues, it can be assumed that they are independent and they all provide different kinds of depth information (Aubin & Arguin, 2014; Landy et al., 1995; Young et al., 1993). If more than one depth cue is present – which is often the case – some form of interaction is happening. Cues have different weights varying from one image to another. The importance of a depth cue is determined by its reliability compared to other cues that exist in the respective image. Each cue is first processed separately. In a second step, the resulting depth estimations for each cue are combined by assigning weights to each cue according to the reliability of this cue. The respective weights are inversely proportional to the reliability of the respective depth cue (e.g. Ernst & Banks, 2002; Jacobs, 2002; Landy et al., 1995). The method proposed by Lovell et al. (2012) corresponds to the MWF model, but they call it maximum likelihood estimation (MLE). In addition to Landy and Johnston (1995), they suggest that depth cues can be dynamically (re)weighted when there are different cue combinations (Lovell et al., 2012). However, Vuong, Domini and Caudek (2006) point out that the MLE approach as well as the MFW model are also limited and seem not to be useful when complex interactions happen.

Bülthoff and Mallot (1988) suggest four different types of interaction between depth cues: accumulation, cooperation, disambiguation and vetoing. The accumulation of depth cues can occur in form of a linear summation where every depth cue has an individual weight. More cues are supposed to result in more depth perception. Here, the integration happens through weighting and adding the cues (Hubona et al., 1999). The weighted additive interaction is very similar to the weak fusion model of Clark and Yuille (1990, cited in Young et al., 1993).

In other cases, cooperation of depth cues often occurs when cues need to complement each other either because all or some of them are weak cues (Bülthoff & Mallot, 1988). Hubona et al. (1999) calls this interaction multiplicative. The cooperation represents a non-linear interaction of cues. Another non-linear interaction is the disambiguation where one cue reduces the ambiguity of another one. Cooperation and disambiguation both are similar to the strong fusion model described by Bülthoff and Mallot (1988).

Bülthoff and Mallot (1988) also found in their studies that stereopsis overrides shading as a depth cue. According to their findings, the depth perception is reduced by 25%, if the two cue conflict in an image (Bülthoff & Mallot, 1988). They concluded from their studies that the interaction of cues proceeded non-linearly through vetoing and inhibition. Such a non-linear interaction can occur through cue promotion according to Landy et al. (1995). In a veto interaction an actually unambiguous cue is sometimes questioned by another one (Bülthoff & Mallot, 1988). It is suggested that stronger depth cues override weaker ones in a conflicting situation in a way that almost only the stronger cue contributes to the depth perception.

Clark and Yuille (1990) were the first to distinguish two approaches: the weak and the strong fusion model (Clark and Yuille, 1990, cited in Young et al., 1993). In the weak fusion model, depth information is processed separately for each cue and afterward linearly combined using different weight (Hubona et al., 1999). In contrast to this, the strong fusion model proposes that cues are not integrated linearly, but in a way that one cue influences another one which leads to depth derivation (Hubona et al., 1999).

All types of interaction are not necessarily mutually exclusive. Bülthoff and Mallot (1988) suggest that performance of participants should increase with more depth cues as there is more information available. This was also observed from Landy and Johnston (1995). However, Vuong et al. (2006) disproved this with showing that cues cooperate rather than simply accumulate and therefore not necessarily lead to more information only by quantity. Vuong, Domini, & Caudek (2006) support a non-linear integration of the cues as they showed that the performance of their participants with a combination of cues is better than expected with a linear model (Vuong et al., 2006). Vuong et al. (2006) compared shading and stereo cues and found – in contrast to Bülthoff and Mallot (1988) – a significantly improved performance when both cues were available. They suppose that the reasons for this might lie in the natural covariation of both cues.

In summary, depth perception has many different components deriving from visual inputs, attention or expectation. Many of them are also related to previous memory or learning (Westheimer, 2011). Depth cues can be interpreted very differently depending on the used image settings and the specific combination of cues. This results in a depth perception that varies a lot depending on the type and quantity of depth cues (Bülthoff & Mallot, 1988).

2.4 Assumptions

Our visual system makes assumptions about circumstances that are also applied to images we look at. These assumptions are often called priors or biases.

One bias is that human beings tend to perceive three-dimensional objects rather as convex than concave (Langer & Bülthoff, 2001; Lovell et al., 2012). Liu and Todd (2004) also found a strong bias for convexity, but they also emphasize that biases are varying individually in their strength.

Another assumption is that we expect one light source that illuminates the objects is placed above us (Cavanagh & Leclerc, 1989; Kleffner & Ramachandran, 1992; Lovell et al., 2012; Mamassian & Goutcher, 2001; Saraf et al., 2005; Sun & Perona, 1998) (**Figure 2.5**). This is also observed by Reichel & Todd, (1990) as their data show that a depth inversion can also occur without shading. This assumption can be justified by the ecological aspect that we experience every day the sun as main light source from above (e.g. Liu & Todd, 2004; Ramachandran, 1988). However, this bias is shifted to the left resulting in a preferred light source between 300° and 330° north-west (Sun & Perona, 1998). Human beings automatically assume that the lighting comes from north-west (Bernabé-Poveda et al., 2005; Saraf et al., 2005, 2007). This left bias was also observed by other researchers, for example by Gerardin et al. (2007) or by Mamassian and Goutcher (2001). Also the cartographic convention for the illumination angle of shaded relief maps lies in the north-western region (e.g. Bernabé-Poveda et al., 2011, 2005). The optimal angle of illumination however is discussed: conventionally the angle is set to 315° whereas Biland (2014) and Biland and Çöltekin (2016) found that 337.5° results in a higher accuracy of land form detection. These findings can be confirmed by Andrews, Aisenberg, D'Avossa, & Sapir (2013) that found mean left biases between 330° and 352°.

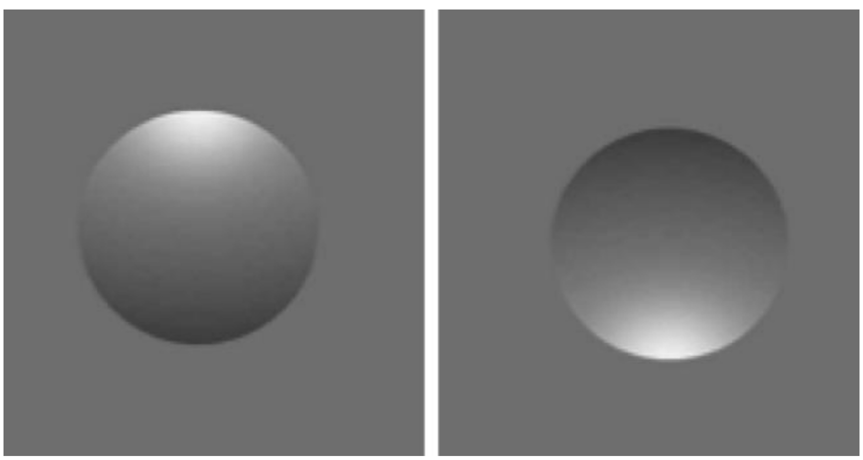


Figure 2.5: The form is shaded brighter on one side and darker on the opposite side in order to imitate the illumination direction. This has a powerful influence on how we perceive the shape of the form (Gerardin et al., 2007).

The reason for the influence of biases is subject to discussion. Sun and Perona (1998) concluded from their studies that the lighting preference has not a biological reason, but is experience-related. This explanation can be confirmed by Adams et al. (2004). They suggest that priors are constantly adapted with interactive experiences with the environment. In contrast to this argumentation, Mamassian and Goutcher (2001) suggest that the light direction bias originates from a preference in the visual field. In a variety of studies, a preference for the left side was observed (Andrews et al., 2013). Andrews et al. (2013) suggest that cultural differences influence the preference of illumination direction. Andrews et al. (2013) conducted a study on different directions of a light source whereby they considered both the handedness of the participant as well as the cultural background. They found a possibly relevant influence of the cultural differences (Andrews et al., 2013). Andrews et al. (2013) also suggest that the left bias might also derive from a hemispheric dominance in our brains.

Although we experience in natural surroundings the sun as the only main light source, this expectation does not represent all everyday situations (Lovell et al., 2012). Often, the direct illumination of the sun is not given and an diffuse distribution of light leads to more complex situations (Lovell et al., 2012; Pentland, 1989). Whenever people look at shaded images – reliefs or satellite images – they assume automatically and inherently that the lighting source is placed above of them (Andrews et al., 2013; Gil et al., 2010). Several researchers suggest that estimating the light source direction from the top is based in a low level mechanism, developing in early visual areas (Andrews et al., 2013; Gerardin et al., 2010; Humphrey et al., 1997; Mamassian, Jentzsch, Bacon, & Schweinberger, 2003). Other authors confirmed this suggestion by showing that the light source is placed mostly in retinal or head-centric coordinates (e.g. Kleffner & Ramachandran, 1992). Langer and Bülthoff (2001) found a 30% larger bias for global convexity than for overhead illumination. This finding is in line with Liu and Todd (2004) who also found a bias for overhead illumination, but it was much smaller than the bias for convexity. However, there difference between the biases was even larger than the one found by Langer and Bülthoff (2001). Liu and Todd (2004) conclude from their studies that the overhead illumination bias is relatively weak and can be overridden by other biases.

The reason to build up these assumption is to allow a stable and consistent perception although depth cues can appear ambiguous (Lovell et al., 2012). In any case of ambiguity in interpreting an image, the human vision system makes assumptions as for example the overhead illumination (Morgenstern, Murray, & Harris, 2011). Not only priors help by deciding which shape or depth we are seeing, also stronger depth cues (e.g. disparity) help disambiguate weaker depth cues (Lovell et al., 2012).

Despite the explanation of the overhead illumination bias is used to explain the terrain reversal effect, Liu and Todd (2004) claim that there are other factors influencing the reversal. They discuss that it is more usual for humans to see surfaces from above rather than from below (global orientation bias) (Baoxia Liu & Todd, 2004). This implies that the depth of a surface increases with the height in an image plane. This finding is in line with Langer and Bülthoff (2001), Mamassian and Landy (1998) and Reichel and Todd (1990). Langer and Bülthoff (2001) found that judgments of participants were based strongly on the perceptual biases of convexity, overhead illumination and global orientation. The observers could not utilize additional information (Langer & Bülthoff, 2001). In contrary to these findings, Liu and Todd (2004) found an improved performance when more information (e.g. about shadows) was present.

2.5 Correction Methods

Correcting the terrain reversal effect is a difficult task as it is highly individual and not yet fully understood when and where the effect occurs. It is difficult to generalize observers or environments (Gil et al., 2014; Liu & Todd, 2004; Morgenstern et al., 2011). Proposed methods in theory rely often on the change of the illumination direction in the original image (Gil et al., 2014; Wu et al., 2013) or on the change of the viewing angle (Zhang et al., 2016). Zhang et al. (2016) provide an useful overview on existing correction techniques. They classify the methods into 'changing the viewing angle', 'indirect change (SRM not included)' and 'direct change (SRM included)'(Zhang et al., 2016). This classification is kept in the following presentation of different correction methods, but is expanded to more methods.

Changing the Viewing Angle

For the remediation of correction methods, different techniques are presented from different authors. One possibility is to rotate the image by 180° (Bernabé-Poveda & Çöltekin, 2014; Bernabé-Poveda et al., 2005; Saraf et al., 1996). This method ties on the fact that the sun's position cannot be changed, therefore the observer's position in relation to the source is adapted. This resolves the terrain reversal effect and the depth is perceived correctly. This method is applicable on low cost and effort and is the easiest way to remove the effect (Wu et al., 2013). Also, it is applicable on analogue images (Saraf et al., 2007) and preserves color information (Zhang et al., 2016). However, this method results in an inversion of the northern direction (**Figure 2.6**). This leads to orientation and interpretation difficulties as changing a map mentally is a challenging task for a human being (Saraf et al., 2007). Additionally, other maps or images have to be rotated in the same way if compared to a corrected satellite image (Saraf et al., 2007). In addition, Liu and Todd (2004) found that cast shadows could also influence the perception of terrain reversal. They found that visible cast shadows make the surfaces more resistant against terrain reversal – but also against a correction – when the image is flipped by 180° (Baoxia Liu & Todd, 2004). An important disadvantage of this method is that the reversion of 180° adds the illusion when there is actually none in the original image (Bernabé-Poveda & Çöltekin, 2014).

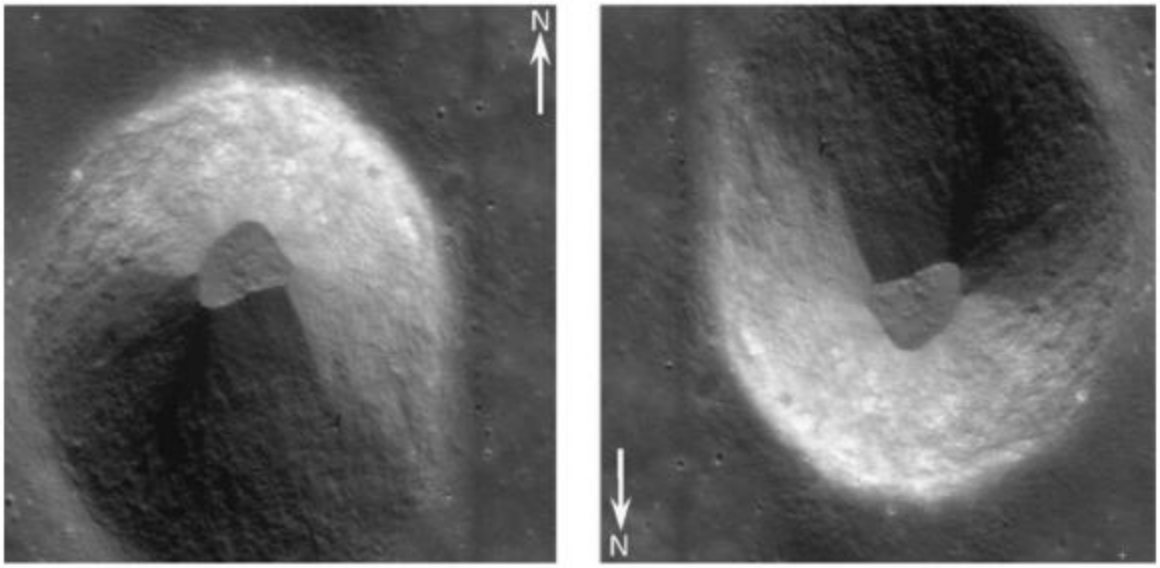


Figure 2.6: Example for the effect of rotating an image. On the left, north is directing towards the top of the image as usual in satellite images or reliefs. On the right side, the image is 180° rotated and north is facing towards the bottom (Wu et al., 2013).

Indirect Change

Another correction can be achieved by taking the negative of an image (Bernabé-Poveda et al., 2005; Saraf et al., 1996). This method changes the original DN values to the opposite value by subtracting the original DN value from 255. In this way, original bright will appear dark and vice versa. This method has the advantage that the north direction remains the same. However, colors are strongly distorted and terrain reversal remains if valleys and ridges point in different directions (Saraf et al., 1996). This method is mostly useful for greyscale images or single band images (Saraf et al., 2007). However, interpretation is difficult due to the reduced information about the surface and texture of objects. Gil et al. (2014) confirm that this method is not useful for color composites due to the alteration of radiometry and low contrast. This statement was confirmed by Bernabé-Poveda et al. (2005) and Bernabé-Poveda et al. (2011). Bernabé-Poveda et al. (2005) point out that semantic discrepancies occur for lakes, paths and vegetation that make interpretation more difficult.

Another method aims to enhance image classification and lineament mapping relies on color-balancing the original image while correcting for the terrain reversion (Saraf et al., 2007). In this method, a false color composited image is transformed from a red-green-blue (RGB) representation to an intensity-hue-saturation (IHS) channels. To correct the terrain, a negative is created to reverse the illumination conditions. Then the image is transformed backwards from IHS to RGB again (Saraf et al., 2007). One restriction is that the brightness of the image can be reduced or overflown and water bodies are not displayed as in the original image (**Figure 2.7**). However, Saraf et al. (2007) claim that the image classification is enhanced with this method. In contrast, Wu et al. (2013) claim that this method cannot process greyscale images.

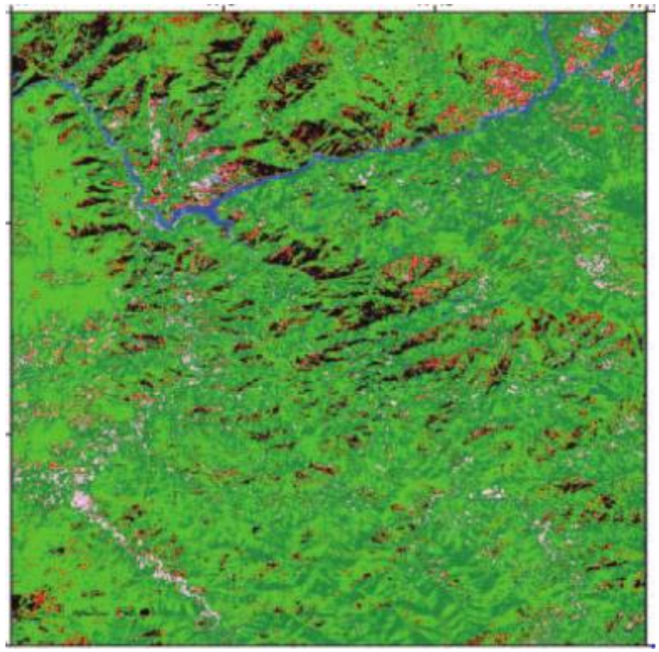


Figure 2.7: Example of a terrain reversal corrected image proposed by Saraf et al. (2007). The colors and the brightness of the image are strongly influenced.

Correction techniques discussed by Bernabé-Poveda et al. (2011) concern the color respectively pixel adjustments. For example, pixel values can be stretched towards white or black, emphasizing the contrasts (Bernabé-Poveda et al., 2011). However, this technique degrades the color proportion strongly. Another correction only processes the brightness and not the hue (Bernabé-Poveda et al., 2011).

Direct Change

Direct changes are subdivided into corrections through fusion and corrections through overlay.

Fusion

Image fusion relies on a technique that applies a SRM to the original satellite image and adapts the image by a weighted coefficient of low-pass filter (Zhang et al., 2016). Different transformation algorithms aim to preserve the color intensity (Zhang et al., 2016). Reviews and overviews of these methods are found by Bernabé-Poveda et al., (2005) as well as Zhang et al. (2016).

Another option is to transform the original satellite image from RGB to IHS. The image is transformed backwards (from RGB to IHS) using the SRM as intensity image and the hue and saturation images from the original satellite image to receive a terrain-corrected image. This is a robust and convincing method with a realistic looking result (Saraf et al., 2005). However, Saraf et al. (2007) claim that it is time consuming. In contrast to these arguments, Gil et al. (2010) state that this method of correcting is more efficient in image processing and it does not rely on stereo pairs. Additionally, the availability of digital elevation models (DEM) is a prerequisite. However, a SRM can also be extracted from digitizes contour lines (e.g. Gil et al. 2010). Also, it can be difficult to represent fully shadowed areas as this results in unusual colors (Saraf et al., 2005; Saraf et al., 2007). This is confirmed by Bernabé-Poveda et al. (2011) as dark shaded pixels in the SRM will make dark areas in the satellite image even darker and light areas will be lighter. However, such situation are rare and very dark regions can also not be classified completely using the original image (Saraf et al., 2005). Saraf et al. (2005) conclude that the interpretation and classification is enhanced with this correction method. The study of Gil, Armesto and Cañas (2005, cited in Gil et al., 2014) preceded the one of Saraf et al. (2007): they used a principal component analysis from a fused image of a Landsat-Thematic Mapper (TM) composite and SPOT (Satellite pour l'observation de la terre) panchromatic data. They also processed the IHS system and weighted the P-channel of the SPOT image (Gil et al., 2005 cited in Gil et al., 2014).

Wu et al. (2013) also suggest a method that relies on a SRM. However, here the SRM is fused with the original image using a shift invariant wavelet transform (SIDQT) method. Together with IHS and principal component analysis (PCA), the wavelet-transform based method belongs to the image fusion techniques. The method decomposes information and reconstructs them with high precision while keeping color distortion minimal. As it is a shift variant method, it provides stable and consistent fusion results (Wu et al., 2013). This method combines the low-pass filtered intensity with the SRM (Zhang et al., 2016).

Image fusion techniques however result in poor color conditions (Zhang et al., 2016). They might be helpful, but not for the use in remote sensing as the radiometric information is adapted (Gil et al., 2014). This makes subsequent processing or classification of the data difficult.

Overlay

A further possibility is to overlay a satellite image with a semi-transparent SRM (Saraf et al., 2005). With this method, a shaded relief model is applied which uses a sun azimuth angle that opposites the angle of the original satellite image (Saraf et al., 2007). The shading adds information about three-dimensional spatial relationships that are neither observable in the satellite image nor in a DEM (Gil et al., 2014).

While creating the SRM, the sun azimuth angle is changed by 180° to achieve that the original satellite image is perceived differently lit (**Figure 2.8**). Traditionally, a shaded relief model has a azimuth angle of 315° and 45° altitude (Bernabé-Poveda et al., 2011). This convention was also used in the correction method applied by Gil et al. (2010). The selection of the sun azimuth and the sun elevation angle is crucial (Saraf et al., 2005). This is also confirmed by Gil et al. (2014) as they experimented with two different solar incidence angles (45° and 60°). The SRM and the original satellite image possess the same spatial resolution in order to keep the same image size and extent (Saraf et al., 2005). This factor was also considered by Gil et al. (2010) and Gil et al. (2014) as they indicated that the resolution of the DEM should not differ a lot from the resolution of the satellite image as this would cause distortions in the corrected image. This finding was confirmed by Wu et al. (2013). The modification of the sun light direction is convenient, but not trivial as light scattering, radiometric correction and the recording time are important factors (Bernabé-Poveda et al., 2011).

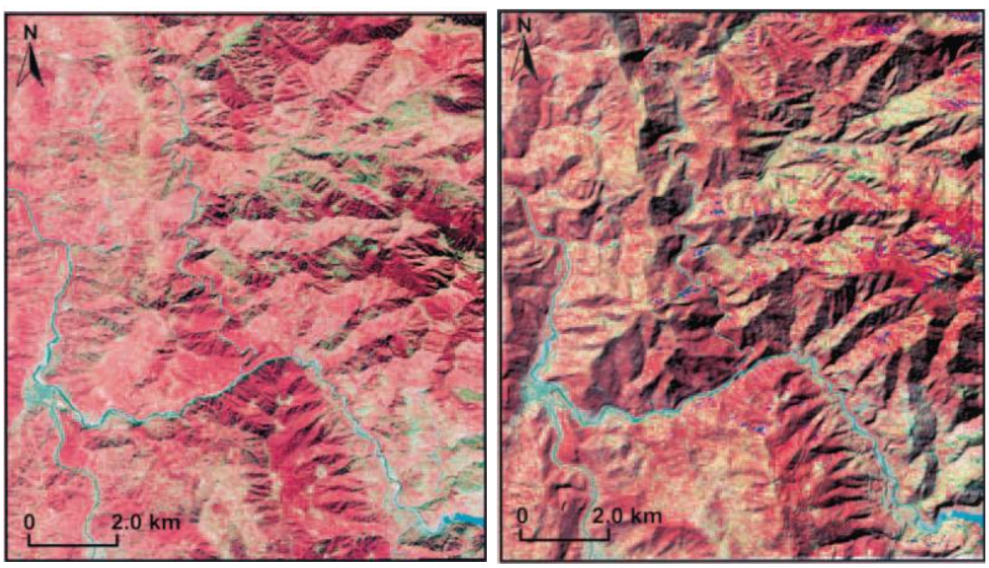


Figure 2.8: Example of the SRM-overlay correction method. A semi-transparent SRM was placed over a false color composite. The left side shows the original image containing the terrain reversal effect (the rivers seem to flow on top of a ridge). The right side shows the corrected image (Saraf et al., 2005).

According to Gil et al. (2014) the SRM overlay method is the most commonly used correction method. Gil et al. (2014) discuss a correction method that relies on the correction presented by Gil et al. (2010). A transparent SRM was put on top of a satellite image with terrain reversion. The opacity level depends a lot on the image types and can vary between 30% for panchromatic images and 50% for multispectral images (Gil et al., 2014). Gil et al. (2014) found that fused images with higher resolution are well suited to overlap with a SRM as they do not lose any identification information. The overlay of an STM has the strong advantage that the corrected image is already orthorectified and oriented towards the north as well as the initial radiometric characteristics are kept (Gil et al., 2014). Therefore, image classification can be applied afterwards without constraints. Additionally, this method is suitable for different kinds of satellite imagery and different acquisition modes as well as in satellite images with different characteristics (Gil et al., 2014). It is also a simple and fast method (Gil et al., 2014). Disadvantages are found in the loss of sharpness due to the overlay and the subsequent color desaturation (Gil et al., 2014). This is especially unfavorable for images with low spatial resolution. Additionally, also the image texture is affected by the superimposing of the SRM (Gil et al., 2014). Gil et al. (2014) point out that using an SRM for correct shading information is helpful, but only an approximation to the true shading. This might require a constant comparison of the shaded layer and the original satellite image to avoid misinterpretations.

To overcome the color deficiency resulting from the semi-transparent SRM, different techniques can be applied. For example, using a RGBA method, the RGB channels correspond with the values of the SRM, but the alpha channel is fully transparent in flat areas, and only corrects the hilly areas where the terrain reversion happens (Bernabé-Poveda et al., 2011). But after all, colors are still desaturated. For a method that overlays a SRM, Gil et al. (2010) observe also a loss of spectral information. However, they argument that this loss is only visual and it would not affect the classification of the image, but rather complement it (Gil et al., 2010).

A method that is similar to the SRM-overlay of Saraf et al. (2005) is presented by Bernabé-Poveda et al. (2005). This method also tries to enhance color information. As in the method of Saraf et al. (2005), a semi-transparent relief model is overlaid on top of a satellite image. Afterwards, the contrast is adapted to enhance the color information (Bernabé-Poveda et al., 2005).

Other Corrections

Zhang et al. (2016) provide an overview of the different topographic correction approaches that have been developed until now. They all rely on the processing of the color hue. While they try to keep the original color hue, they also often generate extreme brightness and darkness values (Zhang et al., 2016). Another rather simple solution is the replacement of shadows with neutral colors and textures to remediate the effect. However major disadvantages are the poor result and the false texture information (Wu et al., 2013).

2.6 Stereo and Motion

Stereopsis and motion are supposed to be very strong depth cues. Several studies have investigated their effect on depth perception. In the following the effects of stereo and motion are discussed. It is important to mention that only few studies investigated stereopsis and motion in relation with the terrain reversal effect (Meyer, 2015; Willett et al., 2015). Most studies consider general depth perception.

Stereo

Stereoscopic vision let the human beings extract more detailed information about distance and depth of a 2D image. A human being has two forward-facing eyes which perceive two slightly disparate, overlapping images. This little angular difference makes it possible to perceive depth and distance with a third dimension (Westheimer, 2011). Stereoscopic vision provides the observer with more qualitative and quantitative information than a single image (Gil et al., 2010).

A major issue for stereoscopic images is the way of presenting it. The possibilities of displaying stereoscopy were developed a lot since the presentation of a stereoscope by Charles Wheatstone in the year 1832 (Sexton & Surman, 1999). Display systems can be classified into stereoscopic systems that rely on external equipment such as glasses and autostereoscopic systems that do not use any additional aids (Sexton & Surman, 1999). Mehrabi et al. (2013) distinguishes also a real 3D group. Techniques for stereoscopic vision rely on binocular parallax. Therefore, a separate image for each eye is presented with a light shifting. The anaglyph techniques are subset of stereoscopic techniques and use a color-separation method to achieve stereopsis. It is one of the most common methods (Gargantini, Facoetti, & Vitali, 2014). In this method, the observer uses a pair of glasses with different color filters (e.g. red and cyan) for looking at the image that is separated into the components of the stereo pair (Sexton & Surman, 1999). This method of perceiving stereopsis is low-cost and simple to apply. More than one person can see the images at a time and it is also applicable in hard copy images However, the quality of the glasses is crucial and the color information in the image is reduced. Additionally, this method can cause nausea and discomfort and image ghosting can occur when the overlapping of the two images is not optimal (Mehrabi et al., 2013; Řeřábek, Goldmann, Lee, & Ebrahimi, 2011; Westheimer, 2011).

Many previous studies have explored the use of stereo images for depth perception. Hubona et al. (1999) found that using stereo images, participants performed better in terms of response time and accuracy compared to non-stereo viewing. They indicate a dominant influence of the stereo cue in relation to the shading cue. Hubona et al. (1999) say that their results are consistent with the vetoing and strong fusion mechanisms but not with the additive or multiplicative interaction model. Other previous studies investigated the combination of the depth cues ‘shading’ and ‘disparity’ show that the reliability generally is increased and that disparity often outranges shading cues (e.g. Bülthoff & Mallot, 1988). Lovell et al. (2012) for example found that the shading cue was only half as reliable as the depth cue of disparity. Other studies confirmed these findings but also showed that estimations of individuals did not become more accurate (Vuong et al. 2006). Other studies showed that the response time of judgements of depth were shorter with shading and disparity in combination (Schiller et al. 2011; Zhang et al. 2007). However, the resulting reliability of a depth cue is dependent on the individual situation: variations in lighting conditions and the shape and texture of the viewed object influence the reliability strongly (Adams et al., 2004; Lovell et al., 2012).

If stereoscopic images are used, discussion about the ability of seeing in depth arises. Different estimations are made for the ability of human beings to see stereoscopically. Some researchers say that only about 5% of the population have a lack or poor stereo vision (Hess, To, Zhou, Wang, & Cooperstock, 2015; Westheimer, 1994). In contrast, Hess et al. (2015) infer from their studies that 32% have moderate or poor stereo vision. There is no conclusive explanation why almost a third of the population should have reduced stereo vision. However, Hess et al. (2015) suggest that this has neural reasons and might be reversible. The reversibility of stereo vision is also experienced at Sacks (2006) although this relies only on a personal report of one individual.

Stereopsis was found to result in an increased depth perception (Hartle & Wilcox, 2016). Hartle and Wilcox (2016) also found that prior experience with stereoscopic images influenced the performance of participants significantly. Therefore, a learning effect of how to use binocular disparity with other cues can be observed during experience time (Hartle & Wilcox, 2016). Although, stereopsis is widely discussed in depth perception and also compared to other depth cues, the use in context with the terrain reversal effect is rather limited. Meyer (2015) compared in her study images containing the terrain reversal effect in a non-stereoscopic and a stereoscopic display. The results showed that stereoscopy has an influence on the perception of terrain as participants were more accurate with stereo images (Meyer, 2015).

In summary, stereopsis and correction using stereo images is a powerful depth cue providing information about location, size, shape, orientation and depth (Hubona et al., 1999). However, not all studies showed enhanced results for stereo images: Van Beurden, Kuijsters and IJsselsteijn (2010) found no effect of stereo on accuracy and response time of individual’s answers.

Motion

The extraction of shape or depth from motion is not new to perception research. While other cues are widely used, the motion cue is not explored extensively (Willett et al., 2015). The motion cue can be classified into two groups: object motion and motion parallax. For object motion either the object moves (uncontrolled object motion) or is moved by the observer (controlled object motion). For motion parallax, the observer moves itself, respectively the head in order to perceive depth (e.g. Van Beurden et al., 2010).

Object motion

The perception of depth through motion is related to the kinetic depth effect (KDE). This effect shows that people could recover three-dimensional information by viewing two-dimensional projection (Hubona et al., 1999 and Vezzani et al., 2015). They explain that the perception of a third dimension relies on a learned association between two-dimensional projection and a three-dimensional structure (Vezzani et al., 2015). The kinetic depth effect, also called structure from motion (Vezzani, Kramer, & Bressan, 2015), is considered a powerful depth cue. It refers to the perception of depth caused by moving two-dimensional stimuli (Vezzani et al., 2015). In contrast to stereoscopy, the perception of depth through motion is a monocular cue that can create strong impressions of depth (Vezzani et al., 2015). Motion of objects comes in different forms: translation, curl or rotation, uniform divergence or deformation. Only deformation, meaning the contraction in one direction and expansion in the opposite direction will give information about an object's shape (Vezzani et al., 2015). This is also the fundament of the studies from Willett et al., (2015). Willet et al. (2015) explores depth and shape information in terrain maps using motion as depth cue. Motion was here possible through interactive relief shearing or animations (Willett et al., 2015). In this relief shearing, low-elevation points remained close to their original position while points on higher elevation were shifted laterally. In the interactive version, the user could grab and drag a point on the map. Different variants were compared: traditional two-dimensional maps, animated shearing that sheared the terrain continuously and integrated shearing where participants interactively sheared and the map (Willett et al., 2015). The results of their study showed that depth perception was increased both with animated and with interactive maps compared to the static map (Willett et al., 2015). Integrated shearing seem to result in more accurate elevation discrimination than shaded relief maps or perspective views (Willett et al., 2015). Between the animated and the integrated variant, very small differences in performance were found. This is in line with findings from Van Beurden et al. (2010): there was no significant difference in accuracy and response time between motion parallax and object motion.

One restriction to the motion cue in case of objection motion is when objects are shallow or not very close to the observer due to perspective reasons (Vezzani et al., 2015). Additionally, motion cues are not automatically unambiguous. Effects such as the terrain reversion can also occur in interactively shearing maps (Willett et al., 2015). However, Willett et al. (2015) experienced less terrain reversals when using interactive variants.

Motion parallax

In contrast to the structure from motion, when using motion parallax the observer moves in relation to the object (Vezzani et al., 2015). This can provide a powerful and unambiguous source of information about depth and structure (Rogers & Graham, 1979). The motion parallax is similar to the structure from motion, but includes human proprioception and information from the motor system (Vezzani et al., 2015). To make this possible, the visual system assumes self-motion rather than that the object moves. Also, the visual system prefers a two-dimensional motion that contains as little movement as possible (Vezzani et al., 2015). Van Beurden et al., (2010) compared object motion, motion parallaxes and stereo in regard of accuracy, completion time, workload and discomfort. They found that both object motion and motion parallax yielded higher accuracy than images without motion. Regarding workload and discomfort, object motion outperforms motion parallax (Van Beurden et al., 2010). Van Beurden et al. (2010) provide an overview over results from different studies comparing stereo, motion parallax and object motion. This overview shows that adding motion or stereo results in an increased performance regarding accuracy and completion time.

Vezzani et al. (2015) point out that only two different two-dimensional images are necessary to recover three-dimensional information about an object. This is similar to the binocular disparity of stereoscopic vision that also uses two images to achieve the perception of a third dimension. Vezzani et al. (2015) propose that, as stereopsis and motion are somehow related in the way they are performed and processed, a tight integration of these cues is expected. Similarities between stereopsis and motion are also reported by Rogers & Graham (1979). The difference lies in the tasks, not in the content: with stereo images, the eyes receive two disparate images simultaneously whereas with motion they see images successively (Rogers & Graham, 1982). Similar to the stereo cue, experience can influence the processing of motion cues (Vezzani et al., 2015). However, also other depth cues are integrated with motion (e.g. Landy & Johnston, 1995).

Both, stereo and motion are considered to be very strong depth cues and are called dominant cues. It was found that motion had a similarly enhancing effect on the task completion time as the use of stereo images (Hubona et al., 1999). This is in line with other findings suggesting that kinetic depth and therefore motion as a cue is equivalent to stereoscopy in terms of performance (Liu & Todd, 2004; Řeřábek et al., 2011; Todd & Norman, 2003). Even more, Hubona et al. (1999) found that the addition of motion was more powerful than stereopsis regarding accuracy. Liu and Todd (2004) suggest that with the combination of stereo and motion, a higher accuracy can be achieved as there are more information available. Mehrabi et al. (2013) even suggest that motion parallax is represent more of a three-dimensionality experience than stereopsis. Also in terms of user’s preference, motion can have advantages as color information is obtained (Řeřábek et al., 2011; Van Beurden et al., 2010). However, motion also includes some disadvantages. As it is a highly complex cue, it might cause difficulties in application and interpretation. Especially when interactivity is added, the observer can be overwhelmed with different kinds of stimuli.

A special form of motion parallax is wiggle stereoscopy presented by Řeřábek et al. (2011). In this method, there is a quick alternation between two slightly disparate images as they are used to create stereoscopic images. This method was compared to anaglyph stereo images and two-dimensional images. The results showed that 3D performed better than 2D. Depth perception of anaglyph stereo and motion parallax is comparable (Řeřábek et al., 2011). In terms of preference, motion is preferred over anaglyph stereo and 2D.

Also recent research discuss the reversion phenomenon as for example by Häkkinen and Gröhn (2016). They describe a reverted waves effect which shows that changes in depth perception not only occur in reliefs but also in material properties of an image (Häkkinen & Gröhn, 2016). This aspect includes much more the surface of objects, in relation to terrain, this would correspond to the land cover. As the land cover is influenced heavily from terrain reversal correction, studies about how surfaces are perceived in combination with depth perception are material for further research. The study of Häkkinen and Gröhn (2016) once again showed the individuality of the inversion phenomenon and therefore points out that basic research of the reversal effect is needed to be subject of research.

2.7 Handedness

Sun and Perona (1998) found values for the left bias in a similar range: they observed a mean left bias of 352° for left-handed and 337° for right-handed participants. However, both are rarely the direction of the sunlight as stated by Bernabé-Poveda et al. (2011). According to Patterson (2016), this convention is in a relationship with a right-handed biased world. One of the earliest explanations for this left bias is delivered by Metzger (1936, cited in Gil et al., 2014). He claims that the asymmetry of preference is attributed to the habit of placing desk lamps on the left side. This is one example of the fact that human beings avoid situations where they have a reduced illumination position for writing or working for example. Sun and Perona (1998) build their hypothesis on this explanation that right-handers would then prefer left-lighting whereas left-handers favor right-sided lighting. A difference between right-handers and left-handers is found, though it is not mirror-symmetric, but both in the same upper left quadrant. The right-hander prefer a stronger left bias than the left-handers (Sun & Perona, 1998). In contrast to this, other studies did not confirm a relationship between handedness and light direction preference (e.g. Andrews, Aisenberg, D'Avossa, & Sapir, 2013; Pascal Mamassian & Goutcher, 2001). For example, Andrews et al. (2013) conducted a study about different directions of a light source whereby they considered both the handedness of the participant as well as the cultural background. In contrast to Sun and Perona (1998), they did not find a significant effect of handedness, but for cultural differences (Andrews et al., 2013).

3 Research Questions

As mentioned earlier, although theoretically discussed, many correction methods for the terrain reversal effect in satellite images are not empirically tested. The correction methods often rely on varying the amount of depth information shown on the satellite images based on depth cues. This master thesis focuses first on comparing different versions of one correction method (SRM-overlay) that appears to be promising based on the literature review (Research Question 1). Secondly, alternative (and some combined) correction methods are applied and empirically tested in a controlled lab study to better assess the accuracy of the tested method and contribution of each depth cue that was featured in the study (Research Question 2). As secondary and exploratory research questions, the influence of handedness is investigated (Research Question 3).

Leading question: Which correction method remediates the terrain reversal effect in satellite images and also preserves the interpretation of them best regarding the performance and subjective experience of participants?

Experiment I:

Research Question 1: Empirically, how well do participants perform with the theoretically promising SRM-overlay correction method in three variants? Specifically, which correction variant is best in terms of accuracy and quality rating for land form and land cover recognition tasks?

Hypothesis 1: It is assumed that with the SRM-overlay method, there is a trade-off between the ability to detect land forms and the perception of land cover. On one side, it is hypothesized that the participants are most effective with a variant providing a compromise between terrain and land cover display.

Experiment II:

Research Question 2: Empirically, are there differences in participants' accuracy, response time, confidence, quality ratings and preferences with the original satellite image, the best variant of the SRM-overlay correction method from Experiment I as well as with three combinations of the SRM-overlay method (adding labels, stereo respectively motion) for land form and land cover recognition tasks?

Hypothesis 2: Correction methods that contain more depth cues than others are expected to perform better in land form detection. It is therefore hypothesized that participants perform better with the combined methods (SRM-overlay with labels, stereo respectively motion) than with the simple SRM-overlay or the original image in land form recognition tasks.

Hypothesis 3: The addition of more depth cues leads primarily to an enhanced land form detection, but does not necessarily influence the land cover perception. It is therefore hypothesized that participants perform equally well within the different correction methods, but better with the original satellite image in land cover recognition tasks.

Research Question 3: Is there a difference between left-handed and right-handed participants regarding their overall accuracy with the different correction methods mentioned in Research Question 2?

This research question is studied in an exploratory way, aiming to propose a hypothesis after the analysis.

4 Methods Experiment I

In the second chapter, several correction methods for the terrain reversal effect were compared. Combining a satellite image with an SRM seems to correct the terrain perception effectively. From all proposed methods, the overlay of a semi-transparent SRM seems to be an effective, robust and feasible correction method using shading as depth cue (Bernabé-Poveda et al., 2005; Saraf et al., 2005; Wu et al., 2013). The major advantages of this method are the improved relief perception, the fact that the north orientation is kept and thematic classifications are still possible although the color perception is qualitatively reduced (Bernabé-Poveda et al., 2011, 2005; Gil et al., 2014; Saraf et al., 2005; Wu et al., 2013). A crucial factor for the success of this correction method is the opacity level of the SRM. In an online study, three different variants of the SRM-overlay method were compared to an unmodified satellite image. The aim is to assess the response time that can be achieved with this correction method as well as the quality rating of the participants for the different visualizations. Also, the best correction variant among the investigated ones is sought.

4.1. Participants

The participants for the online questionnaire were found via email lists and through personal contacts. Participants all over the world were asked with a written invitation via email to take part. No other restrictions were set to the recruiting of participants. A total number of 93 individuals (41 male and 52 female) participated in the study voluntarily.

4.2 Materials

The online study was conducted between February 29th and April 11th. Everybody with a digital device like a computer, laptop, mobile phone or iPad and internet connection could take part in the study. It could not be controlled what the surroundings of the participants looked like and if they took the participation seriously. However, in the case that a person finished the questionnaire completely, the answers were counted as valid. The study took about 25 minutes and was designed using the online tool Survey Monkey[2]. This tool allows the creator to design multiple questionnaires with different types of questions as well as collecting responses and analyzing results. The online study consisted of a pre-questionnaire, a main part and a post-questionnaire (see Appendix for complete questionnaires). On the first page the participants were welcomed and thanked for participating following an introduction containing information on the content and aim of the study as well as indications about data privacy and the length of the study.

[2] https://www.surveymonkey.com/

The participants were always informed in which part of the study they were and how much of the questionnaire they had already completed (in the form of a progress bar).

4.2.3 Post-Questionnaire

After the main part of the study, the participants had to answer some follow-up questions about the study. They first had to indicate how boring/tiring the study was. Afterwards they had to rate the task-wise quality of the used stimuli on a Likert scale ranging from (1) very good to (5) very bad. Another question asked if the participants noticed a contradiction between landform and land cover and if they did, they had to explain these contradictions briefly as well as indicate how they answered the question(s) in this case. In a following section, they were asked if and how often they experienced a switch of landforms (first they saw a valley and shortly after they perceived a ridge in the same image) and how they answered the question(s) in such a case. In the end of the post-questionnaire, the terrain reversal effect was explained and the participants were asked if they had noticed the phenomenon during the study. They were also asked if they have participated in a similar study before. Finally, the participants had the option to leave a comment and/or their e-mail address.

4.2.1 Pre-Questionnaire

The pre-questionnaire contained 16 questions about demographic and personal information of the participant. The questions included information about the participant’s gender, age, level of education, first language and/or English level, writing direction and residence (northern or southern hemisphere or equatorial region). Additionally, they had to rate the frequency of use of the five fields that are relevant for the study (cartography, satellite imagery, 3D-visualizations, photography interpretation and fine arts) on a Likert scale ranging from ‘never’ (1) to ‘very often’ (5). This question was asked first regarding the participant’s professional life and afterwards the same question was asked in respect to his or her leisure time. Subsequently, the participants had to look at a rotating mask and answer a question about the convexity of the face’s features. The participants also had to answer questions about wearing glasses or lenses, color blindness, handedness and hours of sleep.

4.2.2 Main Experiment

After the pre-questionnaire, the participants had to answer 80 questions about land forms and land cover classes. The questions were randomized using the page randomization option in Survey Monkey. The randomization was supposed to avoid a learning effect during the experiment. Before starting the main part, the participants were instructed to answer the questions based on their visual impression and not trying to interpret the images. They also were supposed to answer the questions as quickly, but also as accurately as possible.

4.3 Experimental Design

One of the most promising correction methods that has been theoretically described is the overlay of a semi-transparent relief layer on top of the satellite image (Bernabé-Poveda et al., 2005;. Gil et al., 2014; Saraf et al., 2005; Wu et al., 2013). This method seems to have less disadvantages compared to other correction methods. Therefore, this correction provides the basis for the first experiment. The aim is on the one hand to see which amount of transparency has the best correcting effect. The 'best' correction is the one that leads to the most correct answers compared to the other corrections – it is therefore measured in terms of the achieved accuracy. On the other hand, this first experiment aimed to evaluate how good the 'best' correction variant is. This was again measured in terms of accuracy.

The online study is a 2 x 4 factorial experiment with the factors 'task type' (TRE and LC[3]) and visualization type (Original, Original + 45% transparent SRM-overlay, Original + 65% transparent SRM-overlay, Original + 85% transparent SRM-overlay). Both factors have a within-subject design. Each visualization type contained ten items. Consequently, this comes to 80 items per participant. The participants performed the study on their own without any supervision. Before starting the study, they read an instruction about what they had to do during the study. The participants had the possibility to make breaks as often as they wanted during the study. However, they only could participate once per IP-address.

In the following sections, the correction variants are called:

Original + 45% transparent SRM-overlay = correction 45

Original + 65% transparent SRM-overlay = correction 65

Original + 85% transparent SRM-overlay = correction 85

4.3.1 Independent Variables

In an experiment the independent variables represent the objects of main interest. This variable is manipulated with each manipulation representing a level of the variable. The participant's behavior does not influence the levels of the variable (Martin, 2008).

[3] TRE and LC are abbreviations for terrain reversal effect and land cover.

Stimuli

The stimuli were collected using satellite images and DEMs from the EarthExplorer[4] of USGS[5]. USGS provides free access to a large amount of scientific data with high quality. For this study only satellite images that contained the terrain reversal effect were chosen from three different areas (North America, Canada and China). The choice was subjective and the only requirements for the areas were that they had to be located in the northern hemisphere and that they did not include well known landmarks (e.g. famous mountains). The first requirement is important because the illusion is stronger and more frequent on the northern than on the southern hemisphere due to the tilted position of the Earth (Bernabé-Poveda & Çöltekin, 2014; Saraf et al., 2007). The second requirement covers the familiarity of a certain region or of a well-known landmark. If a participant had prior knowledge about the chosen area, it would affect the results of the study significantly. To prevent this issue, the chosen areas derived from different locations on the northern hemisphere. Also the scale of the images was adapted in a way that no inference to a geographic location was possible. **Table 1** gives an overview over the most important characteristics of the satellite images.

Additionally, the images counterbalanced according to their content. This procedure has also been done by Bernabé-Poveda and Çöltekin (2014), Biland (2014) as well as Meyer (2015). Only one variable changes while all other features remained the same. Half of the images contained convex forms, the other half represented concave forms in the original satellite image. For the land cover the aim was to have a balanced representation of different land cover types. Also, the land form orientation was considered (facing north, northeast, south and southwest). Additionally, all images had the same aerial perspective.

Table 4.1: Characteristics of the Landsat 8 images used in the studies.

Sensor Name	Landsat 8	Landsat 8	Landsat 8
Product Name	L1T	L1T	L1T
Acquisition Date Acquisition Time	2015-08-23 18:24:07	2015-12-02 05:35:08	2015-10-17 18:30:33
Cloud Cover (percent)	0.72	8.23	0.10
Sun Azimuth (degrees)	151.91	161.16	164.17
Sun Elevation (degrees)	49.57	28.29	30.55

[4] http://earthexplorer.usgs.gov/
[5] https://www.usgs.gov/

Shaded relief maps

The DEMs were downloaded using SRTM data from ASTER GDEM V2[6]. ASTER was developed by the NASA and METI and is able to collect pairs of stereo images. These pairs are used to produce 60 x 60 km DEMs. The images are distributed as GeoTIFF files with geographic coordinates (lon, lat) which are referenced to WGS84/EGM96. The second version of ASTER GDEM data (released in 2011) provides an improvement in resolution, a reduction in offset, voids and artefacts as well as a flattening of lakes (Tachikawa et al., 2011). Nevertheless, some artefacts may still exist and therefore, this issue was kept in mind while choosing appropriate areas. The SRTM[7] from 2000 aimed to provide the first near-global set of land elevations. It collected data during eleven days covering 80% of the Earth's land surface between 60° north and 56° south latitudes. The resolution is 1 arc second which corresponds to approximately 30 meters. After downloading the data and opening them in ArcGIS, the projection needed to be transformed from WGS to UTM in order that they had the same projection as the satellite images. Afterwards, the hill shading was created using ArcGIS. For the hill shading a z-factor of 1 was used, the azimuth was set to 315° and the altitude of the light source to 45°. These settings represent the cartographic convention of lighting shaded relief maps (Biland, 2014; M. Gil et al., 2010).

Satellite images

EarthExplorer provides a Landsat archive with L8 OLI/TIRS[8] data. The Landsat 8 satellite was launched in 2013. In a 16 days cycle it collects images with a size of 170 kilometers north-south by 183 km east-west. The pixel size of the multispectral band is 30 meters and for the panchromatic band it is 15 meters (USGS, 2015b). Before downloading an image, EarthExplorer offers the possibility to choose the percentage of cloud cover. This was set to 'less than 10%'. After downloading, Earth Explorer provides additional information about the chosen scene as for example the date and time of acquisition, the cloud cover (percentage), the sun elevation (degrees) and the sun azimuth (degrees). After downloading the Landsat images, they were opened in ArcGIS. The single bands were merged into a 6-band satellite image. To achieve a true color representation, a true color composite was prepared to perceive the image in RGB (4, 3, 2).

---

[6] Abbreviation for: Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM)
[7] Abbreviation for: Shuttle Radar Topography Mission
[8] Abbreviation for: Operational land Imager (OLI) and Thermal Infrared Sensor (TIRS)

Relief overlay correction

To correct the terrain reversal effect in the satellite images, the generated reliefs were applied on top of the satellite images. For this correction method, the amount of transparency in the relief is crucial. As one goal of this first study is to find the best relief-satellite image combination, three different transparency levels were determined and applied to the relief maps. Bernabé-Poveda et al. (2005) exclude 20% and 80% as possible solutions. With only 20% transparency, the satellite image is hardly recognizable and with 80% transparency the relief is not visible in the satellite image. Bernabé-Poveda et al. (2005) suggest a transparency of around 50%. Gil et al. (2010) suggest opacity levels of 50% for panchromatic images,

Figure 4.1: Example compilation (from left to right): original, correction 45, correction 65 and correction 85

30% for multispectral and 45% for fused images. Subjective experience revealed that 30% is too low to achieve an acceptable result. The three selected levels of transparency are 45%, 65% and 85% (**Figure 4.1**). During the selection, it was subjectively considered that the corrections contained both the satellite image and the relief in a visible way and also that there was a clear distinction between the different depictions. Finally, contrast and illumination of the satellite images were very lightly adjusted.

Task Type

The phenomenon of the terrain reversal effect often let the researcher ask only about the land form (Bernabé-Poveda & Çöltekin, 2014; Biland, 2014). However, for the interpretation of a satellite image not only the correct recognition of the land form is essential, but also the correct interpretation of the land cover. Only when each task can be completed, the satellite interpretation task is fulfilled (**Figure 4.2**).

The task type 1[9] asked a perceptual base shape detection question (similar to e.g. Biland, 2014 and Meyer, 2015). The participants had to complete the phrase ‘The line between A, B and C appears as:’ respectively ‘The line between A and B appears as:’. The answers were presented in a Likert scale and ranged from (1) ‘clearly a valley’ over (3) ‘ambiguous’ to (5) ‘clearly a ridge’. It was only possible to give one answer.

[9] In the following called TRE-questions or simply TRE for terrain reversal effect

The task type 2[10] posed a question about the land cover. Therefore, one square was depicted per image onto a subjectively chosen area. Whenever it was possible, an ambiguous area was chosen to place the square. In this way, the recognition of the land cover was not always obvious. Consequently, half of the questions were considered as ‘easy’ (i.e. unambiguous) and the other half was more difficult (i.e. ambiguous). The participants had to answer the question ’What do you see in the red square?’. The choice of answers correlated with what USGS considered as standardized land cover classes (Anderson et al. 1967). LANDSAT satellites provide mostly Level I data. The levels (I-IV) classify data according to the altitude at which the data was gathered (Anderson, Hardy, Roach, Witmer, & Peck, 1976). The levels are transferred to land cover classes and for the Level I data nine classes were listed: Urban or Built-up Land, Ag-

Figure 4.2: Examples for Task 1 (left) and Task 2 (right). On the left participants were asked what land form they recognize. They could choose between (1) clearly a valley, (2), (3) ambiguous, (4), (5) clearly a ridge. On the right, participants were asked what they see in the red square. Possible answers were Forestland, Grassland, Rock/Sand, Snow/Ice, Water, Ambiguous/Not sure and None of the above.

ricultural Land, Rangeland, Forest Land, Water, Wetland, Barren Land, Tundra and Perennial Snow or Ice (Anderson et al. 1967). In this study, we use Level I data on a very small scaled level. Therefore, the list was adapted and reduced to five choices plus two further options. Answer choices were: Forestland, Grassland, Rock/Sand, Snow/Ice, Water, Ambiguous/Not sure, None of the above. It was possible to give more than one answer. The two task types were chosen with the aim of simulating map reading tasks as for example suggested by Phillips et al. (1975). To be as representative as possible, only two task types were chosen (Phillips et al., 1975).

[10] In the following called LC questions or simply LC for land cover

As the terrain reversal effect is a perceptual phenomenon, the terms ‘correct’ and ‘false’ are not suitable. However, for the analysis of the data regarding the performance of the participants a clear definition of ‘correct’ and ‘wrong’ is needed. Therefore, the participants were told that there is no correct or false answer and that they should just answer according to what they recognize. For the analysis of the terrain reversal questions, the software Google Earth[11] was used to get the height above sea level (a.s.l.) of the indicated forms which was compared to the height a.s.l. of a neighboring land form. This is how the correct land form was determined. For the analysis of the land cover questions, two different levels of difficulty needed to be distinguished: in 50% of the images, the participants had to detect one single land cover class (e.g. ‘rock’), in the other 50% of the images they had to detect two land cover classes. In the case of two land cover classes, it was sometimes easier (e.g. ‘snow & ‘water’) to differentiate the classes and sometimes more difficult (e.g. ‘grass’ and ‘forest’). This division should represent the actual situation of satellite images which are not always easy to interpret. They can contain many different land cover classes that are not always easy to distinguish. On most images, it was possible to subjectively decide which land cover class appeared in the square. However, the distinction between forest and grass was always difficult. As an aid to interpret the images correctly, a supervised classification was conducted. The result was compared to the subjective impression and consequently the correct results were determined.

Expertise

In some of the previous studies, expertise had a weak effect (Bernabé-Poveda & Çöltekin, 2014; Biland, 2014; Meyer, 2015) on the performance of a participant. In other studies, there was a significant effect found for expertise (Bernabé-Poveda & Çöltekin, 2014). Experts[12] tend to answer more accurately than non-experts. Therefore, the independent variable of expertise was observed in an exploratory way. The information of experience was used to categorize the participants into two groups: ‘experts’ and ‘non-experts’. Experts are defined as individuals who indicated that they use maps and / or satellite images often (approx. weekly) or very often (approx. daily) in their job as well as participants that use maps and / or satellite images very often in their leisure time. All other participants were categorized as non-experts.

4.3.2 Dependent Variables

The behavior of the participant during an experiment can be measured with dependent variables (Martin 2008). These variables are able to inform about the participant’s performance and about preference or confidence indications.

[11] https://www.google.ch/earth/download/ge/agree.html
[12] Definition of an expert: “a person who has a geographical education at university level, has experience with satellite images and especially works with them frequently” (Meyer, 2015).

Accuracy

The accuracy is considered as the percentage of correct answers. For the TRE-questions, the participants had either answered correctly (1 point) or not (0 points). For the LC-questions they answered the question either completely correct, meaning they chose only the correct answer(s). In this case, the question was considered as *correct* (1 point). There were also partially correct answers possible, for example when one correct and one false answer were chosen. These questions were considered as *half-correct* (0.5 points). In the case that the question was answered completely false or with 'Ambiguous/Not sure', the question was considered as *false* (0 points). The sum of points for each participant was calculated and divided by the total number of possible correct answers. Each question had to be answered to continue the survey. Therefore, all participants answered all questions.

Confidence

For the TRE-questions confidence was measured with the used Likert scale (**Figure 4.3**). The Likert scale is not only a measure for magnitude (here e.g. magnitude of correct answers), but also for confidence (Maurer & Pierce 1998). The Likert scale ranges from 1 to 5 where 1 and 5 are considered as very confident, 2 and 4 as little confident and 3 as not confident. A confidence score was calculated for the analysis. This score contained three levels each of them representing another level of confidence: 1 and 5 (= 'clearly a valley' or 'clearly a ridge') are given two confidence points (= very confident), 2 and 4 got one confidence point and 3 received zero confidence points (= not confident at all). For the LC-questions, only two options were possible: either the participant chose the answer 'ambiguous / not sure' (= 0 confidence points) or another answer (= 2 confidence points). All these other answers were rated as 'confident' as guessing could neither been identified nor excluded. That is why no further distinctions (i.e. between score 1 and 2) could be made. The scores were summed up for each participant and divided by the number of questions, which resulted in a mean score ranging between 0 and 2.

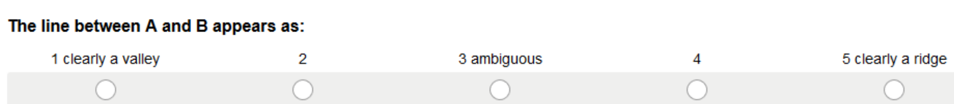


Figure 4.3: Example of the Likert-scale that can be used to determine confidence.

Quality

In the post-questionnaire, participants were asked to rate the quality of the images they saw during the study. The question was asked per task type and for all visualization types. They were asked: ‘To identify land forms (hills, valleys) / land cover (e.g. vegetation, snow, rock): how do you rate the quality of the images above?’. Possible answers were listed on a Likert-scale from (1) very good to (5) very bad. To calculate a quality score, the values have been reversed in order that a higher score represents a higher quality rating as this leads to a more intuitive interpretation of the bar charts and scores. The scores were summed up per participants and divided by the number of questions.

4.4 Statistics

The analysis of the results for both experiments was performed using IBM SPSS Statistics 21[13]. For the theoretical background and the statistical decisions, the book “Discovering Statistics using SPSS” from Field (2009) as well as the homepage of the UZH Methodenberatung[14] were used.

Concerning the descriptive statistics, the mean (M) and standard error (SE) are reported for the parametric tests. For the non-parametric test, the median (Mdn) was indicated. Principally the results are reported as suggested by Field (2009). Results with a p-value of less than 0.05 were considered as statistically significant. Significant results were further investigated with post-hoc tests. Depending on the test different values are reported: F-value (F), t-value (t), degrees of freedom (df), p-value (p), effect size (r), sample size (n).

For preprocessing and visualizing the data, Microsoft Excel 2016[15] was used. The visualizations represent bar charts with indicated error bars. The error bars represent the variation of plus and minus one standard error from the mean. It represents how well the sample represents the population (Field 2009). If the standard error is large in comparison to the sample mean, a larger variability is suspected between the means of the different samples. Thus, such samples might not represent the population very well (Field 2009). This should be remembered in the interpretation of the results.

[13] https://www.ibm.com/marketplace/cloud/statistical-analysis-and-reporting/us/en-us
[14] http://www.methodenberatung.uzh.ch/datenanalyse.html
[15] https://products.office.com/de-ch/excel

5 Results Experiment I

The following chapter presents the results from the online study of the Experiment I. In a first section, the results from the pre-questionnaire and from the post-questionnaire are shown. The main results of the online study are presented afterwards. The effect size was calculated for significant and not-significant results. The interpretation followed Cohen's criteria that r = 0.1 is a small, r = 0.3 a medium and r = 0.5 is a large effect (Field, 2009).

In the following section, land form tasks ('is it a valley or a ridge?') are named TRE-questions and land cover tasks ('what do you see in the red square?') are called LC-questions.

5.1 Participants

93 individuals, whereof 41 are men and 52 are women, participated in the online study. The age ranges from 18 to 79 plus one participant over 80. The majority of the participants – 57% – are between 18 and 29 years old. The participants come from all over the world and have different educational backgrounds. 82 % of them have a university degree. Most of them (89%) do not speak English as their first language. 96% of theses non-native speakers rate their English level between intermediate and proficiency. The spatial distribution of all participants is shown in **Figure 5.1**.

Figure 5.1: Distribution of all participants of the online study. 91% live on the northern hemisphere most of their time. (Source of map: http://www.darrinward.com/lat-long/?id=2209975)

On average, the participants needed 39 minutes to complete the whole study. However, no time constraint was set and the participants could do breaks as long and as many as they wanted. If the participants who needed more than 60 minutes are excluded, the average time for completion is 29 minutes.

Level of Experience

The participants indicated their level of experience with five different fields: cartography (maps), satellite imagery, 3D (geo-)visualizations (e.g. Google Earth, Street View), photography interpretation and fine arts. They had to tell their experience in their job (**Figure 5.2**) as well as in their leisure time (**Figure 5.3**). Generally, the participants encounter maps, satellite images, 3D (geo)visualizations, the interpretation of photography and fine arts more often in their leisure time than for their job. In the job, 40% of the participants work with maps, 27% experience satellite images and 27% use 3D (geo-)visualizations often or very often. More than half of the participants never or rarely use photography interpretation (56%) or fine arts (83%) when working. In the leisure time, 56% experience maps, 27% use satellite images and 35% use 3D (geo-)visualizations often or very often. Photography interpretation (48%) and fine arts (53%) are used never or rarely. In summary, the experience of the participants with different geographic field is balanced out both for the professional field and the leisure time.

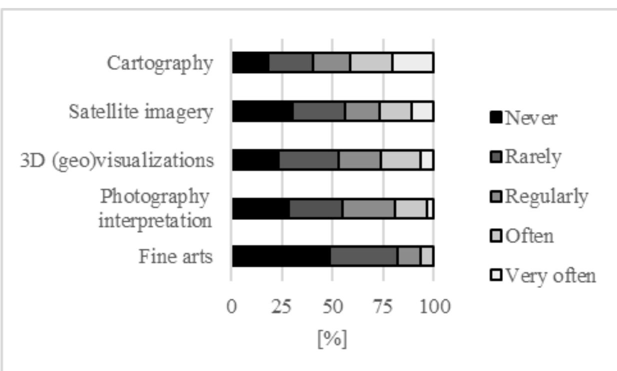


Figure 5.2: Level of experience in **profession / job** for the five different fields. Experience is shown in %. Rarely is approx. once a year, regularly means approx. once a month, often is approx. once a week and very often is approx. daily.

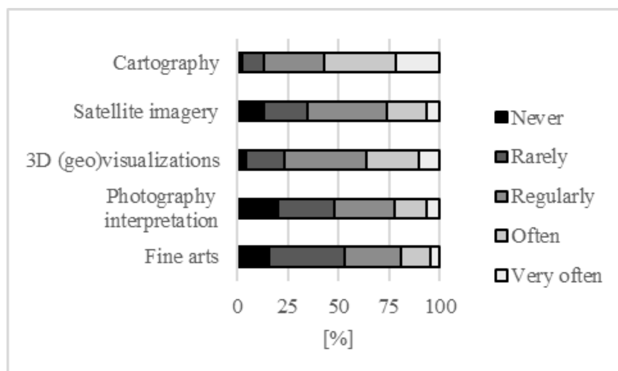


Figure 5.3: Level of experience in **leisure time** for the five different fields. Experience is shown in %. Rarely is approx. once a year, regularly means approx. once a month, often is approx. once a week and very often is approx. daily.

Hollow mask

In the pre-questionnaire, the participants saw a video of a rotating hollow mask. After watching the video, the individuals had to judge if the nose of the rotating mask always appeared convex whenever the nose of the mask pointed directly at the participant. Alternatively, they could indicate that the nose of the mask sometimes appeared convex and sometimes concave. They were told to look at the nose for answering the question. This distinction was made in order that the participants knew where to look. In 76% of all cases the question was answered with ‘yes, the nose always appears convex’ whereas 24% said that the nose sometimes appeared convex and sometimes concave (**Figure 5.4**).

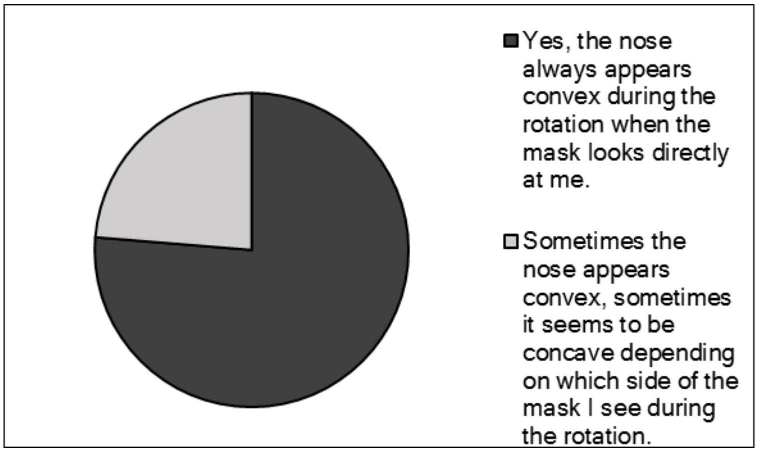


Figure 5.4: Result of the hollow mask task. Most participants indicated that the nose always appeared convex.

Terrain reversal

In the post-questionnaire, the participants had to indicate if they noticed a contradiction between land forms and land cover. The answers were balanced: 48% of the participants indicated that they realized a difference whereas 52% did not notice any contradictions. Some participants indicated snow in a valley as a contradiction or more general that the vegetation looked ‘like in a valley’. Others mentioned that rivers or roads seem to be on ridges. Some respondents noticed a contradiction between shadow and light. In the case that the participants noticed a contradiction, 51% of the participants answered the question based on their perception, 37% replied based on their interpretation and 12% could not remember how they answered the question(s).

At the end of the study, the terrain reversal effect was explained and the participants were asked if they had realized this illusion. From all participants, 52% said ‘yes’ and 48% answered with ‘no’. Also, eight of the 93 participants took part in a similar study before.

5.2 Main Results

Accuracy

For the online study, the accuracy of the answers from the participants were calculated for each visualization type. **Figure 5.5** shows the overall accuracy [%] of the participants' answers. Very little difference is shown among the visualizations: the smallest mean occurring at the original image (M = 58.90, SE = 1.59) and the largest at the correction 65 (M = 66.96, SE = 1.21). In order to extract more information about the correction effect of a visualization, the results were separated by task type for further analysis (**Figure 5.6**).

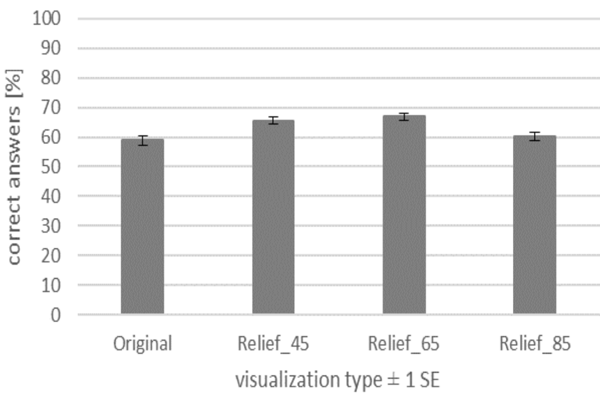


Figure 5.5: Overall accuracy [%] for each visualization type.

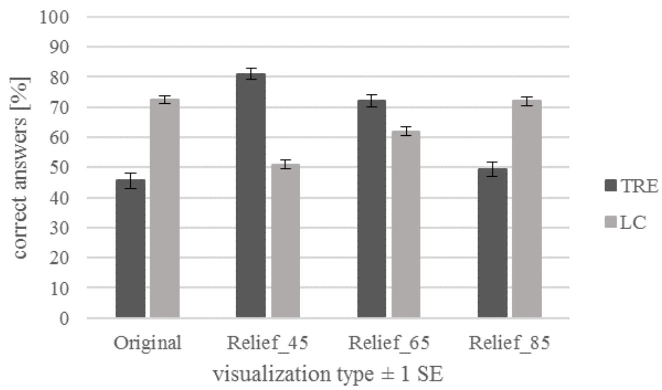


Figure 5.6: Accuracy [%] for each visualization type. The results are separated by TRE-questions (dark) and LC-questions (light). Significant results are not indicated due to lack of space.

For the *TRE-question*, a Friedman test showed that the accuracy of the participants' responses (n = 93) differed significantly among the original satellite image (M = 45.59, SE = 2.65), the correction 45 (M = 81.05, SE = 1.93), the correction 65 (M = 72.15, SE = 1.92) and the correction 85 (M = 49.35, SE = 2.32), $\chi^2(3) = 100.28$, p = .000. Wilcoxon tests were used to follow up this finding. Each correction was compared to the original: The accuracy was significantly higher with the correction 45 (*Mdn* = 90.00) than with the original (*Mdn* = 40.00), z = -6.69, p = .000, r = .69. There is also a significant effect for the comparison of the correction 65 correction (*Mdn* = 70.00) and the original (*Mdn* = 40.00), z = -5.77, p = .000, r = .60. However, with the correction 85 (*Mdn* = 50.00), participants did not answer significantly better than with the original (*Mdn* = 40.00), z = -2.56, p = .01, r = .27.

The accuracy of the participants with TRE-questions was also analyzed among the correction variants. Participants answered significantly less correct with the correction 65 (*Mdn* = 70.00) than with the correction 45 (*Mdn* = 90.00), z = -4.97, p = .000, r = .52. The same significant effect was found when the correction 85 (*Mdn* = 50.00) was compared to the correction 45 (*Mdn* = 90.00), z = -6.57, p = .000, r = .68. The correction 65 (*Mdn* = 50.00) included significantly more correct answers than the correction 85 (*Mdn* = 50.00), z = -5.54, p = .000, r = .57.

For the *LC-questions*, a Friedman test showed that the accuracy of responses of the participants were significantly different between the original satellite image (*M* = 72.39, *SE* = 1.32), the correction 45 (*M* = 50.92, *SE* = 1.48), the correction 65 (*M* = 61.96, *SE* = 1.67) and the correction 85 (*M* = 71.96, *SE* = 1.36), $\chi^2(3)$ = 163.69, p = .000. Wilcoxon tests were used as post-hoc tests. The accuracy was significantly lower with the correction 45 visualization (*Mdn* = 50.00) than with the original (*Mdn* = 75.00), z = -8.09 p = .000, r = .84. Compared to the original (*Mdn* = 75.00), the participants answered also significantly less accurate with the correction 65 (*Mdn* = 65.00), z = -6.44, p = .000, r = .67. The participants answered only with the correction 85 (*Mdn* = 75.00) not significantly less accurate than with the original (*Mdn* = 75.00), z = -0.70, p = .48, r = .07.

Significant differences in participants' accuracy with LC-questions were found among the correction variants as well. The accuracy was significantly higher with the correction 85 (*Mdn* = 75.00) than with the correction 45 (*Mdn* = 50.00), z = -8.08, p = .000, r = .84. The same was true for the comparison between correction 85 (*Mdn* = 75.00) and correction 65 (*Mdn* = 65.00), z = -6.17, p = .000, r = .64. Participants answered significantly more correct with the correction 65 (*Mdn* = 65.00) than with the correction 45 (*Mdn* = 50.00), z = -6.60, p = .000, r = .68.

Quality

In the post-questionnaire, the participants had to rate the visualization types separated by task type. Quali-

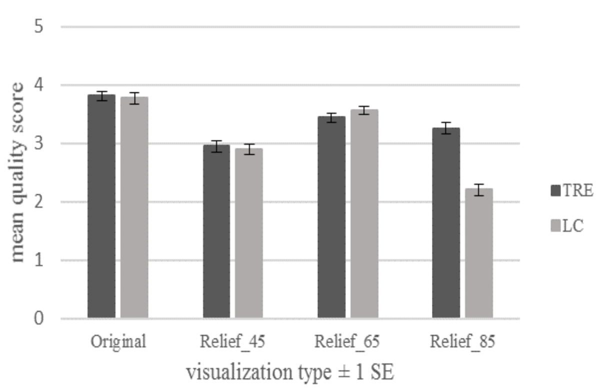


Figure 5.7: Mean quality score for each visualization type and separated by TRE-question (dark) and LC-question (light). Quality scores have no unit and rank from 1 (very bad) to 5 (very good).

ty scores range from 1 (very bad) to 5 (very good) (**Figure 5.7**).

For TRE-questions, the quality was rated best for the original (*M* = 3.82, *SE* = 0.08) and worst for the correction 45 (*M* = 2.96, *SE* = 0.10). For the LC-questions, the quality was rated best for the original (*M* = 3.77, *SE* = 0.10) and worst for the correction 85 (*M* = 2.2, *SE* = 0.09).

Expertise

Expertise is inferred from the experience of the participant with the five different geographic fields (maps, satellite images, 3D (geo-)visualizations, photography interpretation and fine arts) presented in section 5.1, Figure 5.2 and Figure 5.3. Four of the 93 participants (n = 89) had to be excluded from the analysis of the experience as they did not completely fill out the question. An expert was defined as someone that indicated to use maps and satellite images 'often' or 'very often' (i.e. weekly or daily) in the job and 'very often' (i.e. daily) in leisure time. 49 participants were categorized as experts and 40 individuals were considered as non-experts. The accuracy for each visualization type considering expertise and separated by task type were calculated.

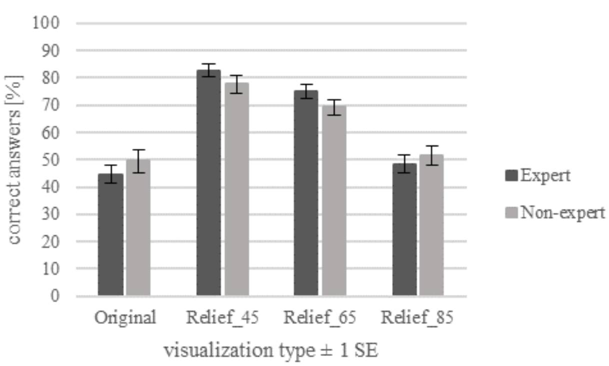


Figure 5.8: Accuracy [%] for **TRE**-questions with each visualization type considering expertise

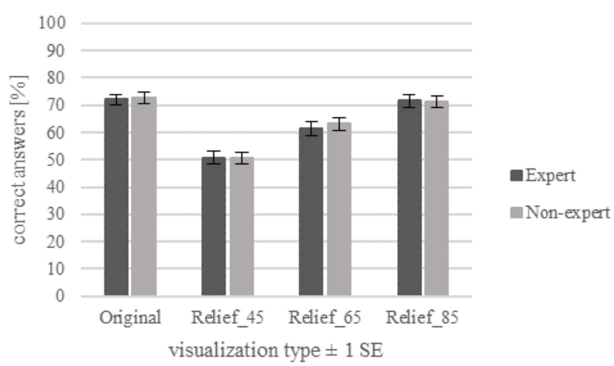


Figure 5.9: Accuracy [%] for **LC**-questions with each visualization type considering expertise

For the *TRE-questions* (**Figure 5.8**), experts as well as non-experts answered most accurately when using correction 45 (Experts: *M* = 82.65, *SE* = 2.40; Non-experts: *M* = 77.75, *SE* = 3.36). The original satellite image led to the least accurate answers for experts (*M* = 44.69, *SE* = 3.46) and non-experts (*M* = 49.50, *SE* = 4.23). The experts answered with lower accuracy when using the original satellite image (*M* = 44.69, *SE* = 3.46) compared to the non-experts (*M* = 49.50, *SE* = 4.23). Also when using the correction 85, experts (*M* = 48.36, *SE* = 3.19) performed worse than non-expert (*M* = 51.5, *SE* = 3.67). However, when using correction 45, experts (*M* = 82.65, *SE* = 2.40) answered more correctly than non-experts (*M* =

77.75, *SE* = 3.36). The same is true for using correction 65 where experts (*M* = 75.10, *SE* = 2.67) gave more correct answers than non-experts (*M* = 69.25, *SE* = 2.91).

For the *LC-questions* (**Figure 5.9**), experts (*M* = 72.14, *SE* = 1.96) as well as non-experts (*M* = 72.63, *SE* = 2.08) answered most correctly with the original satellite image. This result is closely followed by the accuracy of experts (*M* = 71.53, *SE* = 2.32) and non-experts (*M* = 71.38, *SE* = 2.10) when using the correction 85. The least accurate responses were given with the correction 45 for both experts (*M* = 50.81, *SE* = 2.53) and non-experts (*M* = 60.62, *SE* = 1.94). Experts (*M* = 61.42, *SE* = 2.63) differed the most from non-experts (*M* = 63.13, *SE* = 2.35) with the correction 65. Differences between the two groups for all visualization types were more or less balanced out for the LC-questions.

Overall, more differences between the two expertise groups can be observed at the TRE-questions than at the LC-questions.

6 Discussion Experiment I

In this chapter, the results of the previous chapter are analyzed according to the research question 1 developed in Chapter 2. The results are summarized and compared to current research literature.

6.1 Research Question 1: SRM-overlay

***Research Question 1**: Empirically, how well do participants perform with the theoretically promising SRM-overlay correction method in three variants? Specifically, which correction variant is best in terms of accuracy and quality rating for land form and land cover recognition tasks?*

The participants were asked to answer land form and land cover recognition tasks with either the original image or one of three variants of the correction method that laid a semi-transparent SRM over a satellite image. The transparency was set to three different levels: 45%, 65% and 85%. The correction variants are named as followed:

45% transparency overlay = correction 45

65% transparency overlay = correction 65

85% transparency overlay = correction 85

Accuracy

The results showed that if the overall accuracy aggregated over both tasks and for each visualization type was reported, very little difference was seen in the data. The mean accuracy values lie between 59% and 67%. These results, especially the 59%, shows that the accuracy level lies only a little over 50%. Such a result might be interpreted as rather poor performance as an accuracy of 50% can be regarded as success by chance (Liu & Todd, 2004). However, the low accuracy values in this study are rather due to the opposing requirements of the two tasks (land form and land cover recognition). In fact, a correction in one direction helps the performance for one task type while reducing the performance for the other. In other words, the accuracy of answers is found to be higher with a lower amount of transparency in the SRM-overlay. In contrast, the SRM-overlay reduces color and texture information, and therefore a lower level of transparency obscures more land cover information from the original satellite image. Therefore, an SRM-overlay with low transparency results in a reduced amount of correct answers with the land cover identification tasks. Because an even amount of land form and land cover questions were posed in the study, when averaged, overall accuracy appears to be near 50%.

Thus, in order to properly extract the relevant information about the accuracy distribution within a visualization type and between all visualizations, the results were separated by task type. For the land form questions (TRE-questions), a significant difference between the visualization types is found. Applying corrections 45 and 65, participants answered significantly more accurate than with correction 85. Also, participants were more accurate with correction 45 than with correction 65. Correction 45 resulted in an increase of 77% in accuracy compared to the original satellite image in the land form tasks, whereas correction 65 resulted in an increase of 58%. With correction 85, participants answered only 8 % more accurate than with the original. These results show that correction 45 and correction 65 are more suitable for land form detection than correction 85 and the original image. Both corrections receive accuracy values over 70%. As expected, the maximum accuracy is found with correction 65. However, it does not reach 100% (it is slightly more than 80%). This can potentially be explained through the individuality of the terrain reversal effect. The results for the TRE-question show how strongly the overhead illumination bias is guiding out perception (e.g. Kleffner & Ramachandran, 1992; Lovell et al., 2012; Mamassian & Goutcher, 2001; Saraf et al., 2005). As soon as the image is lit from the 'correct', i.e. expected direction, participants perceived more often the correct land form. These results are in line with the theoretical suggestions of Gil et al. (2014) and Saraf et al. (2005).

For the land cover question (LC-questions), participants answered significantly different with several visualization variants. With the original and correction 85, the participants performed best in comparison to correction 45 and 65. This is an expected result as well. However, the best accuracy rate is 72%, achieved with the original image. Participants performed only marginally worse with correction 85 (not even 1% in difference). This shows that correction 85 did not influence the interpretation of the land cover negatively. It also shows that participants had difficulties judging the land cover even with the original as they answered only ¾ of the questions correctly. A reason could be that both experts and non-experts took part in the study. Another reason might be the choice of the satellite images. The quality of the images differs little over all visualizations – even in the original. Correction 45 reduces the accuracy by 42% – almost half of the amount that changes from original to correction 45 with the TRE-questions but inversely. Correction 65 results in a mean accuracy that is reduced only by 17%.

Overall, differences of accuracy within the LC-questions are little smaller than within TRE-questions. In a satellite interpretation it is equally important to recognize the land forms as well as the land cover. The results show that the participants' accuracy with solving both tasks are inversely proportional. Correction 85 shows almost no change in accuracy compared to the original and can therefore be excluded as a correction variant. Correction 45 and 65 both let the participant answer more accurately in total. However, the gap between the mean accuracy of TRE-questions and LC-questions was much larger for correction

45 than for correction 65. Correction 45 can therefore not be rated as a useful correction method, as the mean accuracy value for LC-questions is only 51% and therefore close to guessing the answer. It can be concluded that the best correction variant is correction 65. In literature, several different opacity levels are discussed (Bernabé-Poveda et al., 2005a; Gil et al., 2014). Bernabé-Poveda et al. (2005) and Gil et al. (2014) suggest that the opacity level is adapted for each individual image. For a general correction method, this would be difficult to implement in practice. Therefore, in this thesis, we look for a solution that works for different images although this might reduce the optimal fit in single images. However, the suggestion that a SRM of 65% transparency is a good overlay has to be applied with caution as differently processed images (e.g. different spatial resolution) might need different opacity levels. This is in line with the suggestion of Gil et al. (2014).

Two important factors influence the representation of this correction method and therefore also its accuracy. First, the spatial resolution of the DEM that is used to create the SRM is directly related to the quality of the image (Gil et al., 2014). Second, the opacity levels that were chosen rely on a subjective decision. Gil et al. (2014) and Bernabé-Poveda et al. (2005) suggest different values ranging from 30% to 50%. However, it is difficult to determine an optimal value as the opacity level needs to be adapted regarding image type, terrain, observer and application purpose (Gil et al., 2014).

Expertise

Expertise was investigated as a post-hoc evaluation in respect to the findings of Meyer (2015) who show that experts might answer differently than non-experts. However, these findings show no significant difference. Nevertheless, they are verified in this experiment. Dividing the results for accuracy according to task type and expertise, respectively experience, shows a similar pattern as for the accuracy regarding task type alone. Both experts and non-experts answered the TRE-questions best with correction 45 and the LC-questions with the original and correction 65. Overall, there is little difference between experts and non-experts, especially for the LC-questions. This result is in line with the findings of Meyer (2015). In contrast, the results do not completely accord with the findings of Biland (2014). He suggests that a frequent use of satellite images causes a reduced susceptibility for the terrain reversal effect. Here, experts answered less correctly with the original and correction 65 than non-experts for the TRE-questions. This indicates that also experienced users are influenced by the overhead illumination and the light-from-above-left bias. This argument agrees with the findings of Zhang et al. (2016) who found that the terrain reversal effect also affects experts. It does not coincide with the suggestion that prior knowledge and familiarity reduce the individual’s susceptibility to the terrain reversal effect. Biland (2014) found that only little more than 50% of the experts perceive the terrain correctly in original satellite images. Here it was found, that only 45% of the terrains were answered correctly by experts. The difference might lie in different classification schemes for expertise. Interestingly, experts answered more accurately with strongly corrected images than non-experts. Maybe, non-experts were distracted by the unusual look of the images while experts could here rely on familiarity with different kinds of maps.

Quality

Interestingly, participants rate the quality of the images best for the original satellite image – for both the TRE- and the LC-questions. Although a similar rating would be expected for correction 85, the LC-questions are rated worst with this type of visualization. One reason might be the order of the questions and answers in which they were posed (original – correction 45 – correction 65 – correction 85). The result for correction 65 speaks against this argument as this variant is rated with second best quality for both task types. Correction 65 is rated only 10% worse for TRE-questions and only 6% worse for the LC-questions compared to the original. This is not corresponding to the accuracy achieved with both visualizations. Furthermore, the results show that for TRE-questions, the original was rated best and correction 45 worst. In contrast, the accuracies were inverted to this. In terms of the LC-questions, the original was rated best, but correction 85 was rated worst, while the accuracy of both was very similar. This implies that participants' quality experience and their performance do not correspond. Even if people performed worse, they give good quality ratings and vice versa.

Summary

In summary, the original satellite image and correction 85 perform well for LC-questions in terms of accuracy and quality rating but poorly for TRE-questions. The only correction variant that achieved acceptable results in accuracy for both task types, for experts and non-experts and for the quality rating is correction 65. Therefore, this correction variant is selected to build the base for the second experiment. This is in line with the first hypothesis which expected that the participants are most effective with a variant providing a compromise between terrain and land cover display. The hypothesis can therefore be accepted.

7 Methods Experiment II

The analysis of the first experiment revealed that a SRM overlay with 65% transparency is the best suited correction method for perceiving land forms and land correctly compared to other transparency levels. However, it is not clear if this correction method already yields the best result that can be achieved with such a method or if it can be enhanced. Therefore, a second experiment was conducted and it is the main experiment of this thesis. Two studies were part of this experiment.

The first study compared the original satellite image to the SRM overlay from the Experiment I and to versions of this SRM overlay that were enhanced with labels, stereo and motion. Labels are no natural or pictorial depth cues. They can be considered as artificial cues and have the potential to help perceiving depth (Kruijff & Ii, 2010; Beyang Liu, Gould, & Koller, 2010; Polys, Kim, & Bowman, 2005; Uratani, Machida, Kiyokawa, & Takemura, 2005). However, their role for correcting the terrain reversal effect is not empirically evaluated. Stereo and motion are strong, sometimes dominant cues (Hubona et al., 1999). The implementation of stereoscopic images for the terrain reversal effect revealed significant more accurate results in the study of Meyer (2015). In this case, anaglyph stereoscopy is applied. The motion cue is explored for depth perception, but only rarely for the terrain reversal effect correction (e.g. Řeřábek et al., 2011; Willett et al., 2015). Here, it is applied in form of an animated object rotation. These three extensions represent three different ways of adding more depth information. They are combined with the satellite image that is already corrected from terrain reversal effect with an SRM-overlay. The aim is to enhance depth and probably also land cover information. This way of combining several depth cues an expecting better results refers to a linear additive cue integration in accordance with the findings of Bülthoff, Bülthoff, & Sinha (1998) and Landy et al. (1995). Finally, five different visualizations are compared in a controlled laboratory study and the accuracy, response time, confidence as well as the quality and preference rating of non-experts (see section 7.1) were evaluated. An additional investigation is the difference between right-and left handed participants. Until now, there is no evidence found that handedness influences the perception of satellite images. Nevertheless, the terrain reversal effect is heavily influenced by individual traits (Gil et al., 2014; Liu & Todd, 2004; Morgenstern et al., 2011). Sun & Perona (1998) showed that handedness influenced performance with artificially lit geometric forms. These findings indicate that an effect of handedness might be found in relief maps where the illumination direction is adaptable. There is no evidence that this effect could also occur when shaded relief maps and satellites are combined. Thus, the study explores if there are any indications for a possible influence of handedness.

The second study was conducted parallel to the first study. This study ties on the findings of Biland (2014) and Sun and Perona (1998). Biland (2014) found that in relief maps the optimal incidence angle of the light for land form detection lies at 337.5°. Sun and Perona (1998) found a difference in light direction preference for geometric figures between right- and left handed individuals. The second study aims to combine these studies and to investigate whether handedness has an influence on the perception of the terrain reversal effect and therefore on its correction.

7.1 Participants

Based on the results of the online study, a controlled user study was conducted at the Institute of Geography of the University of Zurich. Participants were recruited via personal connections and they took part in the study on a voluntary basis. None of the participants that took part in the lab study, completed the online study before. This was a necessary restriction because otherwise they could have benefited from a learning effect and prior knowledge about the study content. A total of 35 individuals (18 male and 17 female) participated which is considered a large[16] sample according to Field (2009). All participants were non-experts in geography. A non-expert was defined beforehand as a person who is not working in a geographical field and without geographical education at university level. Also the non-experts do not work frequently (i.e. daily) with satellite images or maps. On one side, experts in geography tended to perform better in previous studies, however, this higher performance did not differ significantly from non-experts (e.g. Meyer 2015, online study of this experiment). Non-experts could possibly struggle more with perceptual phenomena than experts because they are unfamiliar with geographic visualizations. It is therefore interesting to look at this part of the population (Bernabé-Poveda & Çöltekin, 2014; Biland & Çöltekin, 2016; Meyer, 2015). On the other side, non-experts represent a larger part of the population than experts. The participants were divided into two groups according to their handedness: 20 right-handed and 15 left-handed individuals were recruited.

7.2 Materials

The user study was conducted in the eye movement laboratory (lab) at the Institute of Geography at the University of Zurich. A computer containing an eye tracking apparatus is permanently installed in the lab. Additionally, two desks allow to also use a laptop for a part of the study. For the pre- and post-questionnaires as well as for the stereoscopic vision test, a Lenovo T450s[17] laptop was used. This also meant that the participants had to switch between the laptop and the computer during the session. The lab has no windows and the illumination can be controlled. This ensures that all the participants have the same lighting conditions during the experiment. Overall, the study took 60 minutes. All the questionnaires for the pre-, post- spatial ability- and main-questionnaire were created using Survey Monkey as in the

[16] Samples with n > 30 are considered as large samples as – according to the central limit theorem – the sampling distribution is assumed to be normally distributed (Field 2009, 42).
[17] www.lenovo.com/

online study. The stereoscopic vision was an online test. Contrary to the online study in English, the user study was designed and conducted in German.

7.2.1 Eye Tracker

The lab contains the Tobii TX300 eye tracker[18]. The eye tracker is operated by the software Tobii Studio which allows to include online questionnaires into the study. It is possible to combine several stimuli types from different sources into one recording. Additionally, the computer is equipped with a camera acquiring video and sound during the study. The Tobii Studio provides not only the results of the online questionnaires, but offers also eye gaze data. The eye tracking data was collected because they can give indications about how long participants look at which part of the visualization and additionally give information about cognitive processes that take place (Çöltekin, Heil, Garlandini, & Fabrikant, 2009). This might give insights to design benefits and issues. Also, the additional perceptual and cognitive analysis would help understand better how different information are processed in a human's visual system and brain (Bernabé-Poveda et al., 2011).

7.2.2 Pre-Questionnaire

This pre-questionnaire was similarly designed as the one from the online study. However, certain elements were omitted and therefore, ten questions were included into the questionnaire: gender, age, education, frequency of use of the categories 'Kartographie (Karten)'(*cartography, maps*), 'Satellitenbilder'(*satellite images*), '3D-Geovisualisierungen'(*3D-geovisualizations*) and 'Interpretation von Fotographien oder Bildern'(*interpretation of photographs and paintings*), frequency of use of computers, use of glasses/lenses, color deficiency, interpretation of the rotating mask and hours of sleep.

7.2.3 Pattern Folding Test

As spatial ability can be related to the amount of thinking geometrically, it is closely linked to the percep-

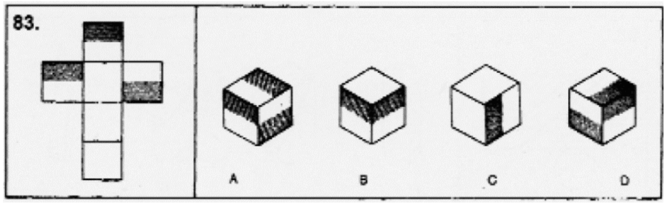


Figure 7.1: Example from the Pattern Folding Test. On the left the open pattern is shown. On the right side, four different figures are shown.

[18] www.tobii.com

tion of three-dimensional structure out of two-dimensional images (Lobben, 2007). Depth perception is highly related to perceiving the third dimension from two-dimensional image and therefore, a pattern folding test was conducted. The pattern folding test aims to investigate the spatial ability of the participants. This test investigates the perceptual ability and the three-dimensional perception (ADA, 2007). In this test a flat pattern with partial shading is presented to the participants along with four different three-dimensional figures (**Figure 7.1**). The participants had to mentally fold the open pattern into a three-dimensional figure and indicate the correct figure. Each time only one answer was correct. The participants had a time limit of six minutes. They were told to solve as many questions as possible, but that they had enough time and did not need to hurry. It was possible to skip a question, however, the participants were told to only skip a question if they were not able to answer a question. The test is composed of totally 15 pattern folding tasks.

7.2.4 Stereoscopic Vision Test

The stereoscopic vision is one of the major advantages for perception and navigation of human beings and animals with front facing eyes. However, not everybody automatically can see stereoscopically. Hess et al. (2015) claim that only 68% of their sample have good or excellent stereoscopic vision, whereas 32% have poor or moderate 3D vision. As stereoscopic images are used in the main experiment, the stereoscopic vision of each participant was tested.

The stereopsis test is online available[19] and it is part of a research study of Prof. Robert Hess and Prof. Jeremy Cooperstock from the McGill University of Montreal. They provide a brief and simple test to assess the quality of a participant's three-dimensional vision which is also discussed by Hess et al. (2015). After the test, a stereo acuity score was calculated and displayed. Stereo acuity the unit of measurements for a person's stereopsis ability. It is the smallest detectable depth difference that someone can perceive (Gargantini et al., 2014). Stereo acuity is defined as "difference in […] two positions converted into an angle of binocular disparity" (Gargantini et al., 2014). This test was chosen as it is a short and freely available test with a scientific background. Additionally, it uses anaglyph glasses – a method that was also used for the stereoscopic images in the main experiment. Before starting the test, the participants had to fill out a handout with information about their height, age and use of glasses/lenses. The monitor size – which is 14 inches as the Lenovo T450s laptop was used – as well as the height of the participant are required information before starting the test. The latter information is used to approximately estimate the viewing distance of the participant from the monitor. To calculate the stereo acuity score, the age of the participants and the use of glasses and/or lenses (answer 'yes' or 'no') had to be indicated. Additionally,

[19] http://3d.mcgill.ca/cbc/

the sitting distance (one arm-length in this case) had to be indicated. The participant had to hold this pre-determined distance for the whole time of the stereopsis test.

For the test anaglyph glasses with amber and blue lenses were used. The test shows different random dot stereograms in which a box appeared that was sometimes behind and sometimes in front of the rest of the image (**Figure 7.2**). The representation of the box was defined by horizontal disparity whereas the rest of the image had no disparity (Hess et al., 2015). The participant had to indicate for each image if the box appeared in front or behind the rest or 'not sure'. The detection of the box became more difficult with the progression of the test. This increase in difficulty was created using eleven disparity levels, ranging from eleven to one pixels (Hess et al., 2015). Before starting the actual test, two example images were shown to the participants to make sure that they realized the difference between the box appearing in front and in the back. Also the participants had time to get used to the anaglyph glasses. After they had understood the examples, they could start the test without any time constraint. During the test one image was shown at a time and after clicking the answer a new image appeared automatically. The test was finished as soon as one of the following situations occurred: first incorrect answer, two consecutive 'not sure' answers or the correct answer at the most difficult stimulus (Hess et al., 2015). A normal stereo acuity score lies between 10 arc seconds and 50 arc seconds depending on the type of test. For this test, the absolute scores were less important than the comparison within all participants. The main interest was to decide if a participant has stereoscopic vision or not.

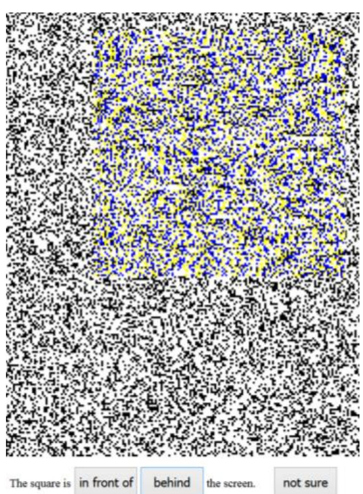


Figure 7.2: Example stimulus from the Stereoscopic Vision Test. Participants had to indicate if the square was in front of or behind the screen or if they were not sure.

7.2.5 Main Experiment

After the prequestionnaire and both ability tests, the main experiment was conducted on the fix-installed computer. First, the eye tracker was activated and the participant was instructed not to touch or move the computer and he/she was informed about the eye-tracking. They were told not to move during the experiment as well as to click only on the answers and on ‘Weiter’(*Continue*) and not to scroll up or down. The main part of the experiment consisted of seven different sections represented each by a questionnaire which was designed using Survey Monkey. All questionnaires were integrated as web-stimuli blocks into Tobii Studio. The questions within one questionnaire were randomized as well as all questionnaires were put in random order among themselves. This aimed to strengthen the effort to reduce a learning effect. After each of the seven blocks, the participants were told that they finished another section and that they had to be manually directed to the next block by the experiment supervisor. The participants were instructed to answer the questions as fast as possible but without any hurry. Also, they were told to answer as much as possible with their perception, i.e. according to what they see and not following their interpretation of the image content. The participants had to use anaglyph glasses for one of the blocks. The glasses were similar to the ones used for the stereo acuity test (see section 7.2.4), but this time they contained red and cyan glasses.

7.2.6 Post-Questionnaire

The post-questionnaire was similar to the one of the online study and it was also created with Survey Monkey. The participants had to answer 16 questions about the preceded experiment. After indicating how tiring or boring the tasks were, quality questions per task type followed for each visualization type. Afterwards the participants had to rank all visualization types according to his/her preference from 1 (= best visualization) to 5 (= worst visualization). Then, two questions about malaise while seeing stereoscopic anaglyph images or moving images followed. Furthermore, the participants were asked if they had realized a contradiction between land form and land cover and if they did, they were asked to quickly describe the contradictions and indicate how they answered the question in such a case. They were also asked about the occurrence and frequency of a land form switch within one image as well as how they answered the question in this case. Afterwards, the terrain reversal effect was described and represented graphically and the question was asked if the participant had realized this phenomenon during the study. Finally, the participants were asked if they had participated in a similar study before as well as they were invited to leave a comment and/or their e-mail address.

7.3 Experimental Design

As mentioned in the introduction to this chapter, the experiment is divided into two separate studies. The two studies were performed together within one experiment sessions and also the same participants conducted both studies. The first study relies to the research questions concerning the correction methods and represents the main study of the experiment. The second study focuses on the terrain reversal in relief maps and the influence of handedness. The first study was designed as a mixed 2 x 2 x 5 factorial design (Martin, 2008). The factors are: the task type with two levels (land form and land cover), handedness with two levels (right- and left-handed) and visualization type with five, respectively seven levels (Original, Original + SRM-overlay from Experiment 1 (correction 65), Original + SRM-overlay + Label, Original + SRM-overlay + Stereo, Original + SRM-overlay + Motion) (Martin 2008). The design is mixed as both a within-subject and a between-subject design is included. The factors visualization type and task type have a within-subject design where all participants are confronted with each level. The factor handedness has a between-subject design as it forms two groups of participants. The second study was designed as a mixed 1 x 2 x 2 factorial design (Martin 2008). The factors are: the task type with one level (land form), handedness with two levels (right- and left-handed) and visualization type with two levels (relief and satellite image corresponding to the relief).

In the following sections, the visualizations are called:

Original

Original + SRM-overlay from Experiment 1 (correction 65) = SRM-overlay

Original + SRM-overlay + Label = Label

Original + SRM-overlay + Stereo = Stereo

Original + SRM-overlay + Motion = Motion

7.3.1 Experiment procedure

Already in an early state of designing the study, a flyer with the most important information about the study and the study procedure was distributed via e-mail or given personally to recruit participants. As all the participants were not member of the geographic faculty. After the study was designed, a few pilot tests were conducted. These tests helped practicing the different sequences of the study. They also raised awareness of the quality of the visualizations and reminded to be verbally precise when giving instructions. Also the time limit could be tested. After the pilot tests were completed, a Doodle link was sent via e-mail to potential participants, together with the consent form and access information to the university.

In the e-mail the participants were asked to read the consent form, sign it and bring it along to the study. This should save time, however, it was also possible to read and sign the consent form just before starting the experiment. The study sessions were conducted between May 20th and June 21st.

The study procedure (**Figure 7.3**) followed a written protocol (see Appendix). Before the participants arrived, different preparations were necessary. The computer and the laptop were turned on and on the laptop, the pre- and postquestionnaire as well as the pattern folding and the stereoscopic vision test were prepared. For each test and for the two questionnaires a personal participant code was marked down on the first page. The code is unique and assures the anonymity of the participant. On the fix-installed computer the Tobii Studio software was prepared, the resolution settings were checked and the keyboard was put aside. The illumination of the lab was set to maximal brightness. The study and all instructions were given in German. After the participants arrived, they were welcomed and asked to hand in the consent form. If they did not have the consent form with them, they received a copy. They were asked if they had any questions about the content of the consent form and if not, they were asked to fill out an information sheet for the stereoscopic vision test. This information sheet contained instructions for the stereoscopic vision test but also had space to indicate information about the participant's height, age and use of glasses or contact lenses. Afterwards, the participants were explained that they would work on the laptop for the first three tests, then switch to the fix-installed computer for the main part of the study and go back to the laptop for the last section. They were told to ask questions if needed any time during the experiment. However, only formal or comprehension questions were answered. Questions about the content were not answered.

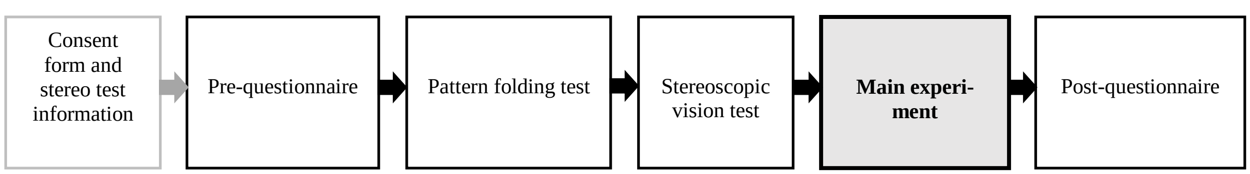


Figure 7.3: Experimental Procedure

The participants then started with the pre-questionnaire. Subsequently, they had to complete the pattern folding test for which they received a short instruction (see section 7.2.3). During the pattern folding test, the time was measured and after six minutes the participants had to stop the test. Before starting the stereoscopic vision test, the participants read the short instruction on the information sheet where they also filled in their personal information in the beginning. While the participants read the information, the experiment supervisor filled in the necessary information for the stereoscopic vision test on the laptop. Afterwards, the participants were positioned correctly (one arm-length away from the laptop) and they were given anaglyph glasses with amber and blue glasses. Then two example images were shown and after the participants were sure that they understood their task, the stereo test was started. After the test, the participants were asked to switch to the computer with the eye-tracker. They received information about the eye-tracking and the procedure of the main part which consisted of seven blocks. They were also advised not to touch the computer and to sit as comfortably as possible because they should not move their head or shoulder a lot after calibrating. They were also asked to only click on the answers and not to scroll during the experiment as well as not to turn their head during or between the different blocks. They also should answer the questions as fast as possible and according to what they see and not to what they would interpret. As the use of anaglyph glasses was needed during one of the seven blocks, the participants were shown where the glasses lie and how they can put them on without moving too much. The participants were also asked to immediately report any indisposition during the study. This was necessary to emphasize as moving images and stereoscopic images were included which might cause malaise. After the instructions, a calibration of the eye-tracking sensor was conducted. After each of the seven blocks, the supervisor had to activate the next block manually. For the block where anaglyph glasses were used, the participants were told that they might need more time to get used to the glasses. After finishing the main part, the participants switched back to the laptop to fill out the post-questionnaire. At the end of the study, the participants were thanked for taking part and a copy of the consent form as well as some chocolate was handed out. The whole experiment lasted around 60 minutes whereby the main part took on average 25 minutes.

7.3.2 Study 1: Correction of the Terrain Reversal Effect

Both studies of the second experiment are conducted as controlled user study. The first study of the second experiment is the main focus of this thesis and treats the different correction methods for the terrain reversal effect. The online study from Experiment I that preceded the controlled user study, examined one promising correction method: the overlay of a semi-transparent relief layer over a satellite image. After the study it was possible to say which variant is best suited to continue with in the controlled user study of the second experiment. Additionally, the online study gave important information about the different task types and how they influenced the analysis of the data (see Chapter 5). The controlled user study used the result of the online study as a base to compare even more corrections. The aim was to investigate if the performance[20] of the solution from the online study can be improved with adding different other correction factors.

7.3.3 Study 1: Independent Variables

Stimuli

The stimuli represented five different visualization types. Two of these five visualizations were inherited from the online study: the original satellite image and the correction with the overlay of a 65% transparent relief layer over the original satellite image. Therefore, the same image selection was used for the controlled user study as for the online study. Otherwise it would not have been possible to compare both results with each other. The correction with the 65% transparent relief overlay provided the basis for the correction methods in the controlled user study. Again, ten items per visualization type were used. This sums up to a total amount of 100 items per participant.

Labelling

The first adaption is created using labels in the satellite image that was corrected with the 65% transparent relief overlay. The addition of labels does not add another depth cue factor. It does not influence the perception of the land form itself, but it adds interpretation information. As humans hardly look at anything without interpreting and drawing conclusions automatically, it can be a solution that might improve the understanding of a satellite image significantly. Furthermore, labelled satellite images are a type of map that humans are used to as almost all freely available map providers offer satellite images with a number

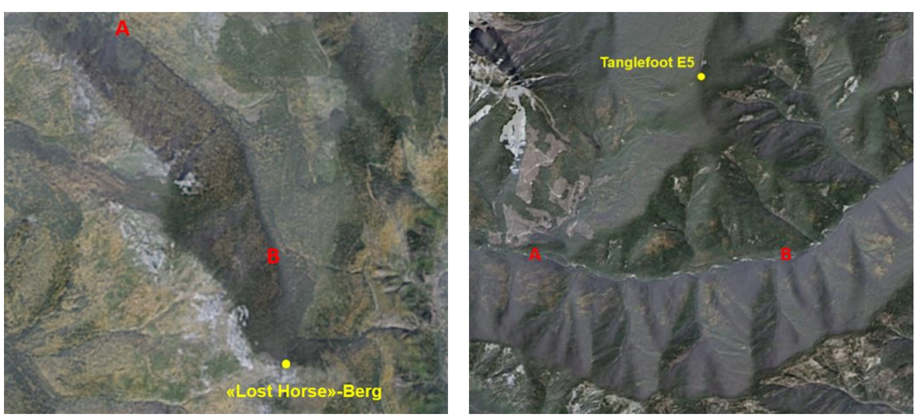


Figure 7.4: Example for the correction using Labels. The left shows an example for an 'easy' example and the right one represents a 'difficult' labelling.

[20] Performance was measured in terms of accuracy in the online study.

of labels. To represent a realistic but comparable situation, only one label per image was introduced. The label was placed as near as possible to the land form that needed to be judged. The labels were also translated into German, in order that no linguistic misunderstandings occurred. The labels were divided into two groups: 50% of the labels were considered as ‘difficult’. This means that they mostly were proper names of a land mark or geographical feature. They did not include any hint to the land form (e.g. ‘hill’, ‘lake’ etc.). The other 50% were considered as ‘easy’ (**Figure 7.4**). They contained at least one reference to a land form. The labels were found on Google Earth. They were displayed with a different color than the markings for the questions (i.e. the A, B, C and the red square). After the question block containing the labelled images, the participants were asked if they had recognized any place names.

Stereopsis

Another visualization was designed using stereopsis onto the relief-corrected satellite image. Meyer (2015) showed that stereopsis can improve the correct perception of land forms although it does not remove the illusion entirely. Stereopsis adds another depth cue factor to the image and influences therefore the perception of land forms directly. The stereo images were created using the software StereoPhoto Maker[21]. Beforehand, two images were created with a small horizontal disparity. These two images were loaded into the StereoPhoto Maker which merges the images automatically. In anaglyph stereo images, ghost effects can occur. Ghost effects prevents the observer to fuse the stereoscopic images (Woods & Rourke 2004). Therefore, the ghost-reduced function of StereoPhoto Maker was used to create the stereo images. Contrast and brightness were adapted minimally. When using red-cyan anaglyph glasses on the

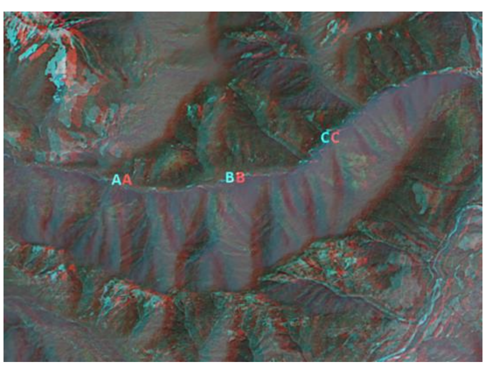


Figure 7.5: Example Stereo Image created with the StereoPhoto Maker. With red-cyan anaglyph glasses, a three-dimensional perception is possible.

[21] http://stereo.jpn.org/eng/stphmkr/

created stereo images, a three-dimensional perception is allowed. For the stereo images, the marking for the questions for each task type (i.e. A, B, C and the squares) were drawn in white color as this enhanced the contrast between them and the rest of the image (**Figure 7.5**).

Motion

The third visualization adds the depth cue element of motion. This potentially adds more depth information directly on the land form perception. It is a strong depth cue that is independent of the shading. The moving images are also created with the software StereoPhoto Maker. Similar as with the stereo images, two images with horizontal disparity were created. In contrast to the creation of stereo images, the disparity had to be reduced to a minimal shift. With the StereoPhoto Maker software, a single image was created that switched between the two images creating a flashing image. The flashing rate was adjusted to a low level (around 10-15 switches x 10 ms) and the x- and y-alignment was adapted in order that the image flashed diagonally from the lower left corner to the upper right corner (**Figure 7.6**). This was the case when x = 8 and y = 4. The result was a moving image that was comfortable to look at. The image was saved as a gif-file.

Figure 7.6: Example from the creation of moving images with StereoPhoto Maker. In the upper left corner, the flashing rate can be adapted. On the lower border the position alignment in the x- and y-

Task Type

As in the online study, the two task types were the land form question (TRE-question) and the land cover question (LC-question). Both task types already are defined and descripted in the section 1.3.1.

Handedness

As mentioned above, the participants were divided into two groups according to their handedness. The crucial factor for the assignment to one of the groups is the hand (right or left) that the participant uses for writing. From the 35 participant, 20 were right-handed and 15 were left-handed. Handedness is expected to matter in the second study that includes reliefs. However, it is also included in the first study as an exploratory approach of observing a difference in the dependent variables with respect to handedness.

7.3.4 Study 1: Dependent Variables

Accuracy and Confidence

For the controlled user study, the definition of accuracy and confidence as well as the calculation process are exactly the same as for the accuracy and confidence in the online study. Both dependent variables are described in section 1.3.2.

Response Time

The Tobii Studio software counts and records every mouse click during the recording. The response time is the time that a participant needed to answer one question. More precisely this is the timespan between clicking on ‘Weiter’ (*Continue*) and the (last) answer choice. For each participant the response time were summed and divided by the number of questions. Response time together with accuracy represent the overall performance of a participant (Martin 2008). Therefore, the response time was also calculated only for the correct answers. Thus, it was possible to compare accuracy with response time only for the correct answers.

Quality and Preference

Both the rating of quality and preference were indicated on a scale from 1 to 5 by the participants. The scale was adapted from the results in order that 1 means low quality or low preference and 5 means high quality or high preference. The preference score referred to each visualization type and was not further processed. The quality score was calculated for each visualization type and with respect to the two task types. For each visualization type and for both task type the mean score was calculated.

7.3.5 Study 2: Influence of Handedness and Left Bias

The second study is attached to the first study and conducted in the same experiment. It serves to answer the side research question regarding handedness. This part of the experiment relies on the research of Sun & Perona (1998) and Biland & Çöltekin (2016). The light direction can be adapted in shaded relief maps whereas in satellite images the light source is given by the relative position of sensor and sun. Biland & Çöltekin (2016) assessed the impact of the illumination direction onto the performance of participants who answered land form tasks. Their results suggest that participants performed better at a light direction of 337.5° north-northwest than at the conventionally chosen 315° north-west. Sun & Perona (1998) found that the preferred lighting is depending on handedness. On average, right-handed individuals preferred a light direction of 336.7° northwest whereas left-handed individuals preferred an illumination direction of 352.1° northwest. The aim of the second study is to explore if performance in land form tasks differed significantly between right- and left-handed individuals.

7.3.6 Study 2: Independent Variables

Stimuli

The stimuli contained several shaded relief maps with different illumination directions. As the preferred light direction lies in the second quadrant (northwest), only illumination direction between 270° and 360° respectively 0° were included. For the stimuli, five different light directions were chosen: 337.5°, 315°, 292.5°, 270° and 0°. Each direction contained eight different shaded relief maps resulting in a total number of 40 items for each participant. All stimuli were created by Julien Biland (2014) and taken over for this study without further processing. In addition, eight satellite images which corresponded to the eight different shaded relief maps were created by using Google Earth (**Figure 7.7**).

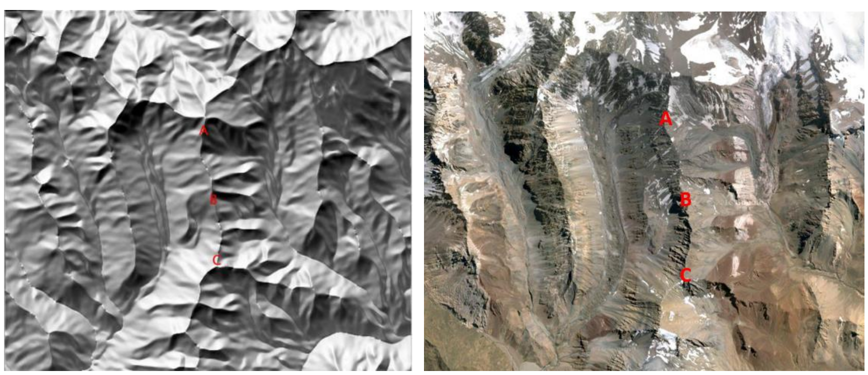


Figure 7.7: Example for the shaded relief map and the corresponding satellite image. On the right side, a shaded relief is shown with an illumination direction of 337.5°. On the left side, the appropriate satellite image is depicted.

Task Type and Handedness

The task for the participants corresponded to the task type 1 that was described in section 4.3.1. The participants had to judge an indicated land form and complete the sentence ‘The line between A, B and C appears as:’. The answer choices were presented on a Likert scale ranging from (1) clearly a valley to (5) clearly a ridge. Concerning the independent variable ‘handedness’, the participants were grouped according to which hand they use for writing.

7.3.7 Study 2. Dependent Variables

Accuracy, Confidence and Response Time

As the task type appears in form of a Likert scale, it is possible to analyze accuracy and confidence at once. These two dependent variables form the inputs for the overall performance of a participant. The response time was again measured with Tobii Studio. The definition and calculation of the three dependent variables coincides with the one from the sections 4.3.2 for accuracy as well as confidence and 7.3.4 for response time.

8 Results Experiment II

The following chapter presents the results from the controlled lab study of the Experiment II. In a first section, the results from the pre-questionnaire and from the post-questionnaire are shown. The results of the main part of the lab study are shown including results for visualization type, task type and handedness and referring to accuracy, confidence, response time, quality and preference.

8.1 Participants

In the controlled lab study, 35 participants thereof 18 men and 17 women took part. Their age ranges from 18 to 79, whereby the majority (60%) are between 18 and 29 years old. All participants live in Switzerland and speak German either as first language or on a very high level. The participants had different educational backgrounds: 14% completed an apprenticeship less than high school (*Lehrabschluss*), 17% finished high school, 9% completed college of higher education (*Fachhochschule*), 23% have a bachelor's degree, 34% have a master's degree and 3% have a doctoral degree. The participants were divided into two groups according to their handedness. The participants could indicate their handedness on a Likert scale from (1) left-handed over (3) ambidextrous to (5) right-handed. For the analysis, the two participants that marked (4) in the Likert scale for handedness, were merged with the right-handed participants. A total of 15 left-handed participants (6 men and 9 women) and 20 right-handed individuals (12 men and 8 women) took part in the study (**Table 8.1**).

Table 8.1: An overview of the two participant groups regarding gender and handedness.

Handedness / Gender	Left	Right	Total
Men	6	9	18
Women	12	8	17
Total	15	20	35

Level of experience

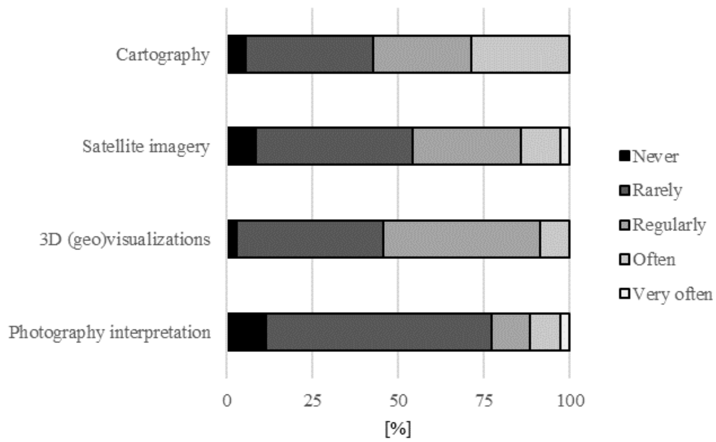


Figure 8.1: Level of experience on the four different fields. Experience is displayed in %. Rarely is approx. once a year, regularly means approx. once a month, often is approx. once a week and very often is approx. daily

All participants were non-experts in any geographic fields. Nevertheless, they were asked to indicate their experience with the fields ‘cartography (map)’, ‘satellite images’, ‘3D (geo)visualizations’ and ‘interpretation of photographs or images’ (**Figure 8.1**). The majority of the answers in each category was answered with ‘rarely’ or ‘regularly’. For cartography the experience was very balanced between rarely (37%), regularly (29%) and often (29%). Satellite images are used rarely from 45% of the participants whereas 31% use them regularly and 14% experiences satellite images often or very often. 43 % of the participants use 3D (geo-) visualizations rarely and 46% use it regularly. More than half of all participants (66%) rarely interpret photographs or images. Overall, the participants hardly chose the ‘extreme’ categories ‘never’ and ‘very often’.

Hollow mask

As in the online study, the participants saw a video of a rotating hollow mask. After watching the video, the individuals had to judge if the nose of the rotating mask always appeared convex whenever the nose of the mask pointed directly at the participant. Alternatively, they could indicate that the nose of the mask sometimes appeared convex and sometimes concave. 60 % of all participants answered ‘Ja, die Nase sieht immer konvex aus’ (*yes, the nose always appears convex*) whereas 40% said ‘Manchmal sieht die Nase konvex aus, manchmal sieht sie konkav aus’ (*sometimes the nose appeared convex and sometimes concave*) (**Figure 8.2**).

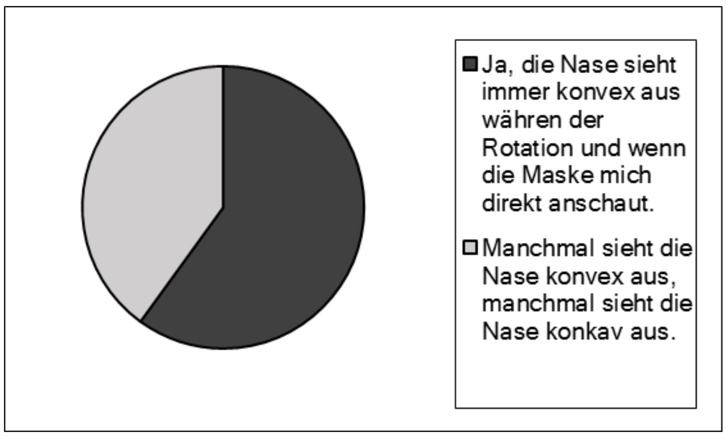


Figure 8.2: Results for the hollow mask task. The majority of the participants indicated that the nose always appeared convex.

Spatial ability

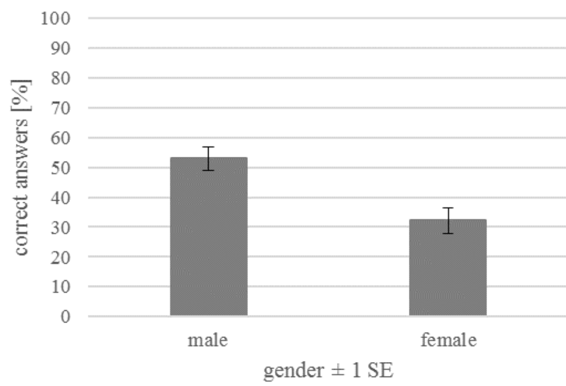


Figure 8.3: Accuracy [%] of the participants in the pattern folding test. The figure shows the accuracy of men and women.

After the pre-questionnaire, the participants had to complete a spatial ability test with a time limit of six minutes. They were asked 15 different questions that all were concerned with the task of pattern folding. Before analyzing the data, the accuracy of each participant in the pattern folding test was calculated. **Figure 8.3** shows the mean accuracy that was achieved in the paper folding test grouped by gender. Overall, male participants gave more correct answers than female participants.

To compare the spatial abilities with the performance of each participant, different correlations were calculated. The performance (in the main part of the study) of the participant is measured primarily by accuracy and response time. The response time was considered only for the correct answers similar to the analysis of Brügger (2015) and Çöltekin, Fabrikant, & Lacayo (2010).These two variables are compared to the pattern folding accuracy. Before each correlation was analyzed, a scatterplot was examined as well as a normality test was conducted for both variables. For normally distributed variables, Pearson's r (r) was calculated as correlation coefficient whereas for non-normally distributed data, Spearman's rho (r_s) was computed. Significant correlations were only found in three comparisons (**Table 8.2**). The accuracy of all answers (meaning TRE- and LC-questions together) with the 'SRM-overlay correction and the pattern folding accuracy correlate significantly in a positive way (r = .37, p = .03, n = 35). The accuracy of all answers given with the 'Label' correction method is also significantly correlated with the pattern folding accuracy (r = .38, p = .03, n = 35). The response time for answering all questions with the original image correlates significantly with the accuracy for the pattern folding test (r = -.36, p = .04, n = 35).

Table 8.2:Overview over all significant correlations. The pattern folding accuracy is compared to the accuracy and response time of the different visualization types with and without separating by task types.

		accuracy SRM-overlay overall	**accuracy Label overall**	**response time Original overall**
accuracy pattern folding	**Pearson Correlation**	.37*	.38*	-.36*
	Sig. (2-tailed)	0.03	0.03	0.04
	N	35	35	35

* Correlation is significant at the 0.05 level.

Stereoscopic vision

After the spatial ability test, the participants did a test that estimated their stereoscopic vision. The test was provided online from the McGill University of Montreal (McGill, 2016). The analysis relies on the paper of Hess et al. (2015) that used the online test for their study.

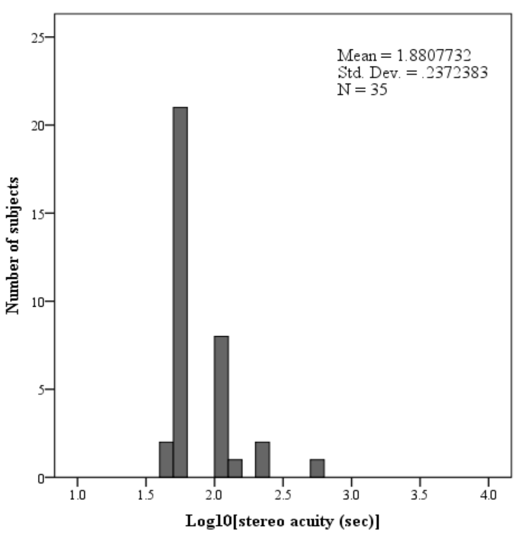


Figure 8.4: Frequency distribution of the stereo acuities [log10(sec)] from the participants.

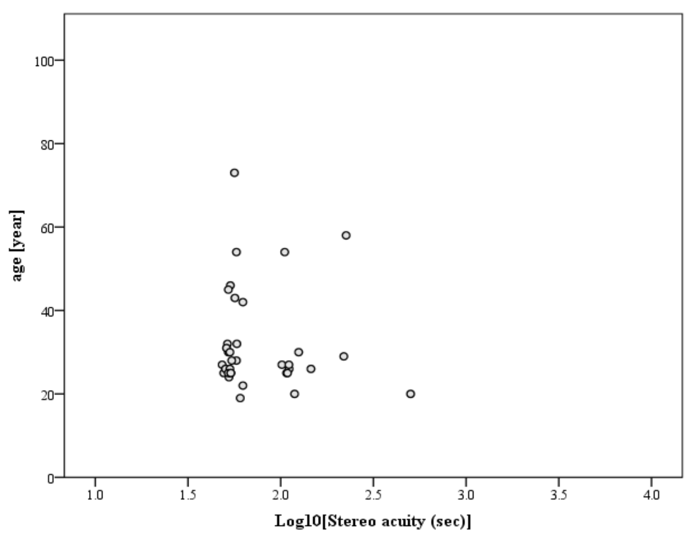


Figure 8.4:The stereo acuity [log10(sec)] in relation to the age [year] of the participants.

The frequency distribution (**Figure 8.4**) shows a peak at 1.88 log arcsec. All values range between 1.68 log arcsec and 2.7 log arcsec. The scatterplot (**Figure 8.5**) represents the stereo acuity in relation to age.

The correlation analysis shows a small negative correlation without a significant effect (r = -.08, p = .65, n = 35). Therefore, the distribution is not age dependent.

The aim of the stereopsis test was to evaluate if the participants could perceive three-dimensional depth at all. Additionally, the result from the stereopsis test was compared to the performance of the participants as a post-hoc evaluation. Performance was measure with accuracy [%] and response time for the correct answer [ms]. The correlation analysis (**Table 8.3**) shows that only weak correlations exist in response time. None of these correlations are significant. Regarding the accuracy, both the overall accuracy and the accuracy related to TRE-questions shows a small to medium sized effect, but no significance. Only the accuracy of stereo-enhanced images with LC-questions and stereo acuity shows a significant correlation (r = -.38, p = .03, n = 35). This is a negative correlation with a strong effect which means that a higher score of stereo acuity correlated with a lower accuracy in ‘Stereo’ images.

Table 8.3: The stereo acuity [log(sec)] compared to the overall accuracy with 'Stereo' correction [%] as well as the ‘Stereo’ accuracy [%] split up by task type (TRE, LC). Further, the stereo acuity was compared to the overall response time [ms] of the correct answers when responding to 'Stereo’ questions. The ‘Stereo’ response time [ms] was separated by task type (TRE, LC) and compared to the stereo acuity.

			Accuracy			**Response time**		
		Stereo acuity	**Overall**	**Stereo_TRE**	**Stereo_LC**	**Overall**	**Stereo_TRE**	**Stereo_LC**
Stereo acuity	**Correlation Coefficient**	1.00	0.06	0.20	-.38*	0.09	0.04	0.06
	Sig. (2-tailed)		0.72	0.25	**0.03**	0.61	0.81	0.72
	N	35	35	35	35	35	35	35

* Correlation is significant at the 0.05 level (2-tailed).

Attention

In the post-questionnaire, the participants had to indicate on a Likert-scale how boring or tiring they experienced the preceding study. They could choose answers ranging between (1) not at all boring / tiring over (3) sometimes boring / tiring to (5) very boring / tiring. Most of the participants (40%) chose (2) which signifies a level of boredom / tiredness between 'not at all' (1) and 'sometimes' (3). 23 % found it not at all boring / tiring and the same percentage perceived the study as sometimes boring / tiring. 11% chose (4) – a level between 'sometimes' and 'very boring / tiring' and only 3% thought that it was very boring / tiring.

Discomfort

The participants were asked if they felt any discomfort while looking at the stereo images with the anaglyph glasses or while looking at the moving images. Again they had to indicate their answers on a Likert scale ranging from (1) no/little discomfort over (3) medium discomfort to (5) strong discomfort. For the stereoscopic images with the anaglyph glasses, the majority (83%) felt no or little discomfort (1) or something between no and medium discomfort (2). 14 % felt medium discomfort and 3% something between medium and strong discomfort (4). For the moving images, the results show the exact same distribution as for the stereoscopic images.

Terrain Reversal Effect

The participants also had to indicate if they noticed a contradiction between land forms and land cover. The answers were balanced: 49% of the participants indicated that they realized a difference whereas 51% did not notice any contradictions. The participants who perceived a contradiction described them as follows: snow was found in a valley and/or near a body of water (28%), a river was perceived on top of a ridge (28%), vegetation did not fit the surrounding (11%) or hills looked like bodies of water (5%). 28 % did not describe a contradiction in their answer. In the case that the participants noticed a contradiction, 65% of the participants answered the question based on their perception and 47% replied based on their interpretation.

At the end of the study, the terrain reversal effect was explained and the participants were asked if they realized this illusion. Out of all participants, 54% said 'yes' and 46% answered with 'no'.

8.2 Main Results

Within the scope of this project, only the results from the first study from Experiment II (concerning the different correction methods) are analyzed (see sections 7.3.2-7.3.4). The results from the online study in Experiment I had shown that the representation of aggregated visualizations is not informative enough to draw conclusions for the usability of a correction method. For that reason, the result from this second experiment are not aggregated for visualization type, task type and handedness, but separated according to the two task types or by handedness.

If a significant effect was observed, post-hoc tests were conducted. These test based on the Bonferroni correction (Field, 2009) if all comparisons were made. The Bonferroni correction aims to reduce type I errors and therefore corrects the alpha value depending on the number of comparisons that are conducted (Field, 2009). Visually, a significant difference is expressed through brackets with an asterisk in the respective figures.

Before performing the inferential analysis, the data are tested for normality using the Shapiro-Wilk test. If the data are normally distributed, parametric tests are performed, if they are not normal, non-parametric alternatives are conducted. Before choosing non-parametric tests for non-normal data, different transformation methods were tested (Field, 2009). These transformations aimed to achieve normality in the data. Also the boxplots were investigated for outliers. If outliers were observed in normally distributed data, the decision whether to keep the values or not was made using an outlier labelling method. If boxplots are used to identify outliers, SPSS uses a step of 1.5 x IQR to determine an outlier. The values that exceed the boundaries of 3 x IQR (in both directions) are considered as extreme values (Tukey 1977, cited in Hoaglin, Iglewicz, & Tukey, 1986). Hoaglin et al. (1986) suggest an alternative method as using 1.5 respectively 3 as multiplication factor is not always an optimal solution. The factor of 3 is a conservative measure and data points lying outside this value can be considered as outliers (Hoaglin et al., 1986). The factor of 1.5 sometimes classifies a value as an outlier although it is not absolutely sure that it is really an outlier. This issue occurs mostly with small samples (Field, 2009). Hoaglin et al. (1986) suggest a factor of 2.2. In order to not judge values too conservatively, this method was applied to evaluate whether or not a value had to be excluded from the analysis. If the data were not normally distributed and no transformation method yielded an effect of normality, the outliers were adapted according to Field (2014). In this case, the scores that were considered as outliers in SPSS were adapted in a way that the score lies one unit above or below the next highest or lowest score in the data set (Field, 2014).

8.2.1 Performance and Confidence – Overview

Before going into a detailed analysis of the correction methods, the data was inspected in order to detect if and to what amount the corrections influenced the participants' answers. For this purpose, the performance (accuracy and response time) as well as the confidence of the participants with the original image was compared to the mean of all correction method together. The aggregated corrections are named 'Corrected' in the figures and they represent the mean value of all four correction methods ('SRM_Overlay, 'Label', 'Stereo' and 'Motion').

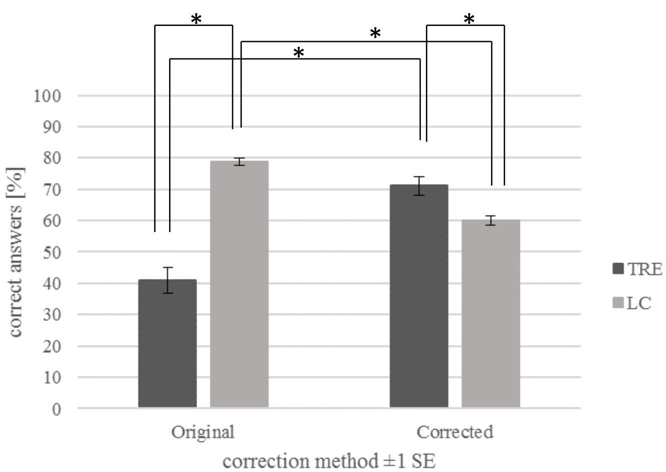


Figure 8.6: Accuracy [%] for the original and the aggregated corrections separated by task type.

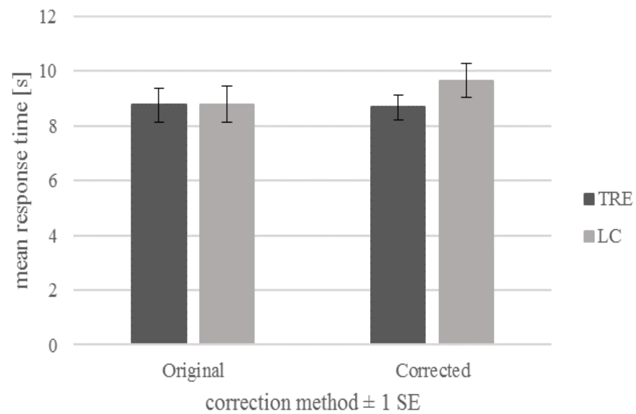


Figure 8.7: Mean response time [s] for the original and the aggregated corrections separated by task type.

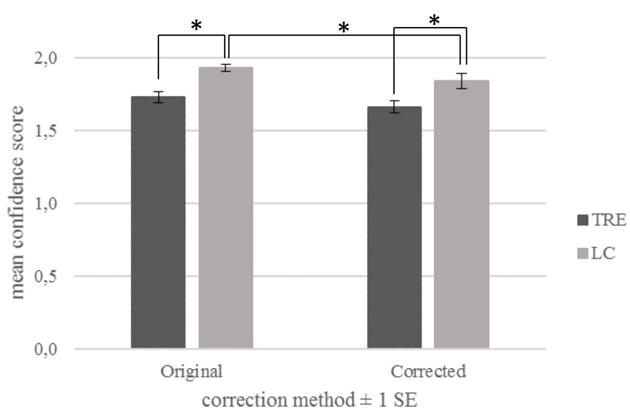


Figure 8.8: Mean confidence score [no unit] for the original and the aggregated corrections separated by task type.

Accuracy

A Friedman test revealed that the participants (n = 35) answered the 'TRE_Original' (*M* = 40.86, *SE* = 4.13), 'TRE_Correction' (*M* = 71.07, *SE* = 2.86), 'LC_Original' (*M* = 78.69, *SE* = 1.16), 'LC_Correction' (*M* = 59.96, *SE* = 1.39), significantly different in terms of accuracy, $\chi^2(3) = 45.28$, *p* = .000 (**Figure 8.6**). A Wilcoxon signed-rank test showed that for the TRE-questions, the original visualization (*Mdn* = 30.00) was answered significantly less accurate than the correction (*Mdn* = 75.00), *z* = -4.28, *p* = .000, *r* = .72. By contrast, for the LC-questions, the answers with the original (*Mdn* = 80.00) were significantly more accurate than with the corrections (*Mdn* = 61.25), *z* = -5.11, *p* = .000, *r* = .86. The accuracy for the original with TRE-questions (*Mdn* = 30.00) did differ significantly from the original with the LC-questions (*Mdn* = 80.00), *z* = -4.90, *p* = .000, *r* = .83. The TRE-questions were answered significantly more correct with the aggregated corrections (*Mdn* = 75.00) than with the aggregated corrections for the LC-questions (*Mdn* = 61.25), *z* = -3.15, *p* = .002, *r* = .53.

Response Time

The response time of the participants was calculated by the Tobii Studio software. We are interested in the performance of a participant, that is composed of the accuracy and the response time. Therefore, only the response time of the correct answers is taken into account to evaluate the performance of a participant. A 2-way repeated-measures ANOVA was conducted. Mauchly's test met the assumption of sphericity as it is a 2 x 2 within-subject design (two task types, two visualizations). Therefore, the degrees of freedom were not corrected. All effects are reported as not significant (**Figure 8.7**). There was a non-significant main effect of the task type on the response time, $F(1, 34) = 1.74$, *p* = .20, *r* = 0.22. There was also a non-significant main effect of the visualization type on the response time, $F(1, 34) = 0.33$, *p* = .57, *r* = 0.10. And finally, there was no significant interaction effect between the visualization type and the task type, $F(1, 34) = 1.95$, *p* = .17, *r* = 0.23.

Confidence

With the Likert-scale for the TRE-questions and the answers for the LC-questions, the confidence of the participants was evaluated. The confidence score lies between 0 (no confidence) and 2 (a lot of confidence). A Friedman test showed that the participants (n = 35) answered the 'TRE_Original' (*M* = 1.73, *SE* = 0.04), 'TRE_Correction' (*M* = 1.66, *SE* = 0.04), 'LC_Original' (*M* = 1.93, *SE* = 0.03), 'LC_Correction' (*M* = 1.84, *SE* = 0.05) significantly different in terms of confidence, $\chi^2(3) = 36.96$, *p* = .000 (Figure 8.8). A Wilcoxon signed-rank test showed that for the TRE-questions, the participants were not significantly more confident when using the corrections (*Mdn* = 1.73) compared to the original visualization (*Mdn* =

1.80), z = -1.86, p = .063, r = .31. But for the LC-questions, the confidence with the original (*Mdn* = 2.00) is significantly higher than with the correction (*Mdn* = 1.95), z = -2.63, p = .008, r = .45. The confidence for the original at TRE-questions (*Mdn* = 1.80) differs significantly from the original at the LC-questions (*Mdn* = 2.00), z = -3.66, p = .000, r = .62. The aggregated correction for the TRE-questions (*Mdn* = 1.73) were answered significantly less confident than the corrections for the LC-questions (*Mdn* = 1.95), z = -3.36, p = .001, r = .57.

8.2.2 Performance and Confidence – Detail

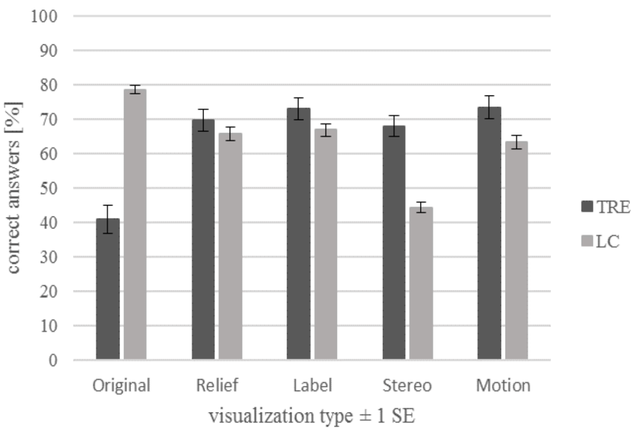


Figure 8.9: Accuracy [%] of different visualization types separated by task type. (Significant results are not indicated due to lack of space.)

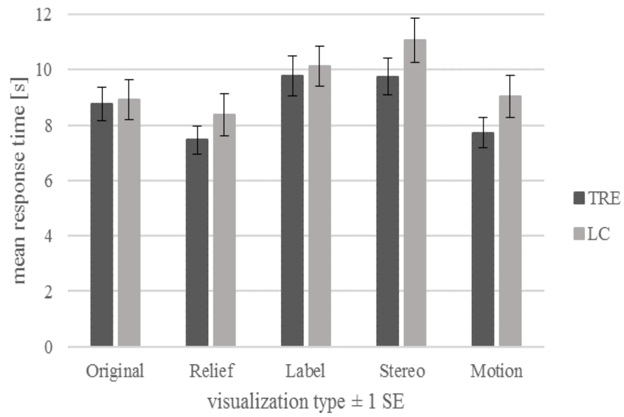


Figure 8.10: Mean response time [s] of different visualization types separated by task type.

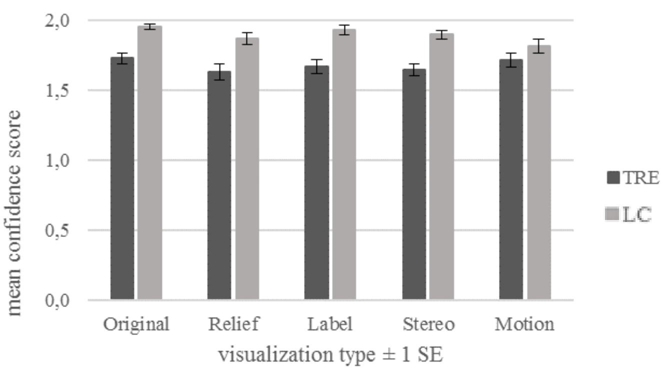


Figure 8.11: Mean confidence score [no unit] of different visualization types separated by task type. (Significant results are not indicated due to lack of space.)

Accuracy

After the general comparison of original satellite images with the aggregated corrections, a more detailed analysis gives insights into the performance and confidence of the participants with individual correction methods.

A Friedman test showed that the participants' (n = 35) answers with'Original_TRE' (*M* = 40.86, *SE* = 4.13), 'SRM-overlay_TRE' (*M* = 69.71, *SE* = 3.29), 'Label_TRE' (*M* = 73.14, *SE* = 3.20), 'Stereo_TRE' (*M* = 68.00, *SE* = 3.06), 'Motion_TRE' (*M* = 73.43, *SE* = 3.38), 'Original_LC' (*M* = 78.54, *SE* = 1.22), 'SRM-overlay_LC' (*M* = 65.71, *SE* = 1.96), 'Label_LC' (*M* = 66.83, *SE* = 1.84), 'Stereo_LC' (*M* = 44.31, *SE* = 1.48), 'Motion_LC' (*M* = 63.29, *SE* = 1.94) are significantly different, $\chi^2(9) = 110.29$, *p* = .000 (**Figure 8.9**).

For the *TRE-questions*, all of the examined accuracies for any corrected visualization are significantly different from the original. With 'SRM-overlay' (*Mdn* = 70.00), participants answered significantly more accurate than with 'Original' (*Mdn* = 30.00), *z* = -4.03, *p* = .000, *r* = .68. With 'Label (*Mdn* = 80.00), the answers are significantly more correct than with 'Original' (*Mdn* = 30.00), *z* = -4.26, *p* = .000, *r* = .72. 'Stereo' (*Mdn* = 70.00) differs significantly from 'Original' (*Mdn* = 30.00), *z* = -3.99, *p* = .000, *r* = .68. The same is true for 'Motion' (*Mdn* = 80.00) which differs significantly from 'Original' (*Mdn* = 5.00)., *z* = -4.29, *p* = .000, *r* = .72.

For the *LC-questions*, also all of the examined accuracies for any corrected visualization are significantly different from the original. With the 'SRM-overlay' (*Mdn* = 70.00), participants answered less correctly than with 'Original' (*Mdn* = 80.00), *z* = -4.32, *p* = .000, *r* = .73. The same applies to the difference between 'Label' (*Mdn* = 65.00) and 'Original' (*Mdn* = 80.00), *z* = -4.11, *p* = .000, *r* = .69. 'Stereo' (*Mdn* = 45.00) differs significantly from 'Original' (*Mdn* = 80.00), *z* = -5.17, *p* = .000, *r* = .87. With 'Motion' (*Mdn* = 70.00), answers are significantly less correct than with 'Original' (*Mdn* = 80.00)., *z* = -4.55, *p* = .000, *r* = .77.

Within the individual correction methods, there are also significant differences between the accuracies for the two *task types*. Two out of five comparisons are significantly different. Answers with 'Original_TRE' (*Mdn* = 30.00) differ significantly from 'Original_LC' (*Mdn* = 80.00), *z* = -4.90, *p* = .000, *r* = .83. Also, participants answered the land form questions significantly more correct ('Stereo_TRE', *Mdn* = 70.00) than the land cover questions ('Stereo_LC', *Mdn* = 45.00), *z* = -4.59, *p* = .000, *r* = .78. The other differences between task types within a correction method are not significant. 'SRM-overlay_TRE' (*Mdn* = 70.00) does not differ significantly from 'SRM-overlay_LC' (*Mdn* = 70.00), *z* = -1.13, *p* = .257, *r* = .19. 'Label_TRE' (*Mdn* = 80.00) does not differ significantly from 'Label_LC' (*Mdn* = 65.00), *z* = -1.67, *p* =

.096, *r* = .28. ‘Motion_TRE’ (*Mdn* = 80.00) does not differ significantly from ‘Motion_LC’ (*Mdn* = 65.00), *z* = -2.62, *p* = .009, *r* = .44.

Response Time

In order to analyze the response time of the participants, a 2-way repeated-measures ANOVA was conducted. Mauchly’s test met the assumption of sphericity for the main effects of visualization type, $\chi 2(9)$ = 12.87, *p* = .17 and task types (N.B. they have only two levels). The assumption of sphericity is also met for the interaction of visualization type and task type, $\chi 2(9)$ = 14.14. Therefore, the degrees of freedom were not corrected and sphericity can be assumed. All effects are reported as not significant (**Figure 8.10**). There was a non-significant main effect of the task type on the response time, *F*(4, 136) = 0.50, *p* = .73, *r* = 0.06. There was also a non-significant main effect of the visualization type on the response time, *F*(1, 34) = 1.31, *p* = .26, *r* = 0.09. There was no significant interaction effect between the accuracy of the participants and the response time, *F*(4, 136) = 0.88, *p* = .48, *r* = 0.06.

Confidence

A Friedman test showed that the participants (n = 35) answered of ‘Original_TRE’ (*M* = 1.73, *SE* = 0.04), ‘SRM-overlay_TRE’ (*M* = 1.63, *SE* = 0.06), ‘Label_TRE’ (*M* = 1.67, *SE* = 0.05), ‘Stereo_TRE’ (*M* = 1.65, *SE* = 0.05), ‘Motion_TRE’ (*M* = 1.71, *SE* = 0.05), ‘Original_LC’ (*M* = 1.95, *SE* = 0.02), ‘SRM-overlay_LC’ (*M* = 1.87, *SE* = 0.04), ‘Label_LC’ (*M* = 1.93, *SE* = 0.03), ‘Stereo_LC’ (*M* = 1.90 *SE* = 0.03), ‘Motion_LC’ (*M* = 1.81, *SE* = 0.05) significantly different regarding confidence, $\chi^2(9)$ = 103.89, *p* = .000 (**Figure 8.11**). Wilcoxon tests were used to follow up this finding.

For the *TRE-questions*, none of the examined confidence scores for any of the corrected visualizations was significantly different from the original. ‘SRM-overlay’ (*Mdn* = 1.80) does not differ significantly from ‘Original’ (*Mdn* = 1.80), *z* = -2.07, *p* = .038, *r* = .35. ‘Label (*Mdn* = 1.70) does not differ significantly from ‘Original’ (*Mdn* = 1.80), *z* = -1.65, *p* = .099, *r* = .28. ‘Stereo (*Mdn* = 1.70) does not differ significantly from ‘Original’ (*Mdn* = 1.80), *z* = -2.15, *p* = .031, *r* = .36. ‘Motion (*Mdn* = 1.80) does not differ significantly from the ‘Original’ (*Mdn* = 1.80)., *z* = -0.08, *p* = .939, *r* = .01.

For the *LC-questions*, one of the examined confidence scores for a corrected visualization was significantly different from the original. Participants answered only with ‘Motion (*Mdn* = 2.00) significantly less confident than with ‘Original’ (*Mdn* = 2.00), *z* = -2.88, *p* = .004, *r* = .49. All other findings for confidence with LC-questions were not significant. ‘SRM-overlay’ (*Mdn* = 2.0) does not differ significantly from ‘Original’ (*Mdn* = 2.00), *z* = -2.24, *p* = .025, *r* = .38. ‘Label (*Mdn* = 2.00) does not differ significantly from ‘Original’ (*Mdn* = 2.00), *z* = -0.36, *p* = .717, *r* = .06. ‘Stereo (*Mdn* = 2.00) does not differ significantly from ‘Original’ (*Mdn* = 2.00), *z* = -1.53, *p* = .126, *r* = .26.

Within the individual correction methods, there are also significant differences between the confidence scores for the two *task types*. Four out of five comparisons are significantly different regarding confidence. ‘Original_TRE’ (*Mdn* = 1.80) differs significantly from ‘Original_LC’ (*Mdn* = 2.00), *z* = -3.86, *p* = .000, *r* = .65. ‘SRM-overlay_TRE’ (*Mdn* = 1.80) differs significantly from ‘SRM-overlay_LC’ (*Mdn* = 2.00), *z* = -3.96, *p* = .000, *r* = .67. ‘Label_TRE’ (*Mdn* = 1.70) differs significantly from ‘Label_LC’ (*Mdn* = 2.00), *z* = -4.31, *p* = .000, *r* = .73. ‘Stereo_TRE’ (*Mdn* = 1.70) differs significantly from ‘Stereo_LC’ (*Mdn* = 2.00), *z* = -4.06, *p* = .000, *r* = .69. Only the confidence scores with ‘Motion_TRE’ (*Mdn* = 1.80) do not differ significantly from the scores achieved with ‘Motion_LC’ (*Mdn* = 2.00), *z* = -1.50, *p* = .134, *r* = .25.

8.2.3 Quality Rating

The participants indicated their preference by judging the different visualization types separately by task type (TRE, LC). This resulted in a preference score ranging from 1 (worst quality) to 5 (best quality).

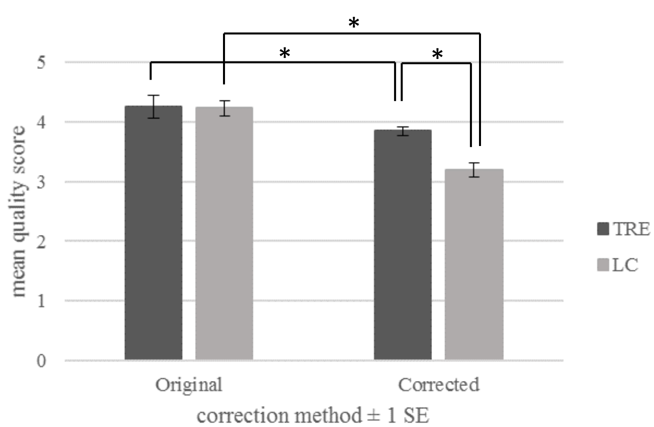


Figure 8.12: Quality rating for the two different visualizations separated by task type.

Accuracy – Overview

Figure 8.12 shows the quality score separated by task type. Here, the original is compared to the mean score for all corrections. A Friedman test showed that the participants (n = 35) rated the quality of ‘TRE_Original’ (*M* = 4.26, *SE* = 0.19), ‘TRE_Correction’ (*M* = 3.84, *SE* = 0.07), ‘LC_Original’ (*M* =

4.23 *SE* = 0.13), ‘LC_Correction’ (*M* = 3.19, *SE* = 0.13) significantly different, $\chi^2(3) = 42.87$, *p* = .000. A Wilcoxon signed-rank test shows that for the TRE-questions, the original visualization (*Mdn* = 5.00) is ranked significantly higher than the corrections (*Mdn* = 3.75), *z* = -2.30, *p* = .02, *r* = .39. For the LC-questions, the original (*Mdn* = 4.00) differs significantly from the corrections (*Mdn* = 3.25), *z* = -4.86, *p* = .000, *r* = .82. The quality rating for the original with TRE-questions (*Mdn* = 5.00) does not differ significantly from the original at the LC-questions (*Mdn* = 4.00), *z* = -0.49, *p* = .62, *r* = .08. However, the aggregated correction for the TRE-questions (*Mdn* = 3.75) were rated significantly higher than the corrections for the LC-questions (*Mdn* = 3.25), *z* = -4.25, *p* = .000, *r* = .72.

Accuracy – Detail

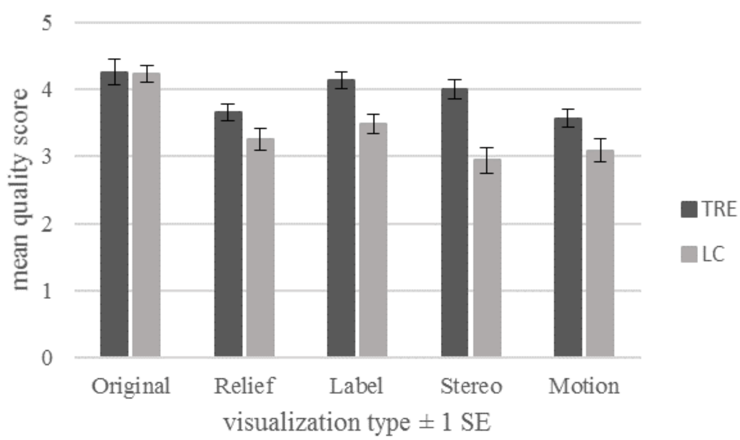


Figure 8.13: Quality for different visualization types. (Significant results are not indicated due to lack of space.)

A Friedman test showed that the participants (n = 35) rated the quality of ‘Original_TRE’ (*M* = 4.26, *SE* = 0.19), ‘SRM_Overlay_TRE’ (*M* = 3.66, *SE* = 0.12), ‘Label_TRE’ (*M* = 4.14, *SE* = 0.12), ‘Stereo_TRE’ (*M* = 4.00, *SE* = 0.14), ‘Motion_TRE’ (*M* = 3.57, *SE* = 0.13), ‘Original_LC’ (*M* = 4.23, *SE* = 0.13), ‘SRM_Overlay_LC’ (*M* = 3.26, *SE* = 0.16), ‘Label_LC’ (*M* = 3.49, *SE* = 0.14), ‘Stereo_LC’ (*M* = 2.94, *SE* = 0.20), ‘Motion_LC’ (*M* = 3.09, *SE* = 0.17) significantly different, $\chi^2(9) = 91.19$, *p* = .000 (**Figure 8.13**). Wilcoxon tests were used to follow up this finding.

For the *TRE-questions*, two of the examined quality ratings for a corrected visualization was significantly different from the original. The quality ratings for ‘Motion’ (*Mdn* = 4.00) is significantly lower than the rating for ‘Original’ (*Mdn* = 5.00)., *z* = -2.71, *p* = .007, *r* = .46. Also, the rating for ‘SRM_Overlay’ (*Mdn* = 4.00) differs significantly from ‘Original’ (*Mdn* = 5.00), *z* = -2.29, *p* = .022, *r* = .39. In contrast, the quality of ‘Label’ (*Mdn* = 4.00) differs not significantly from ‘Original’ (*Mdn* = 5.00), z = -.76, p = .45, r = .13. ‘Stereo’ (*Mdn* = 4.00) differs not significantly from ‘Original’ (*Mdn* = 5.00), z = -1.19, p = .24, r = .20.

For the LC-questions, all of the examined quality ratings for a corrected visualization were significantly different from the original. In terms of quality, ‘SRM_Overlay’ (*Mdn* = 3.00) differs significantly from ‘Original’ (*Mdn* = 4.00), z = -4.35, p = .000, r = .73. ‘Label’ (*Mdn* = 4.00) differs significantly from ‘Original’ (*Mdn* = 4.00), z = -4.25, p = .000, r = .72. ‘Stereo’ (*Mdn* = 3.00) differs significantly from ‘Original’ (*Mdn* = 4.00), z = -4.36, p = .000, r = .74. ‘Motion’ (*Mdn* = 3.00) differs significantly from the ‘Original’ (*Mdn* = 4.00)., z = -4.09, p = .000, r = .69.

Within the individual correction methods, there are also significant differences between the accuracies for the two *task types*. In four out of five comparisons there is a significant difference in quality rating. Only with the original satellite image, there is no significant quality difference between ‘Original_TRE’ (*Mdn* = 5.00) and ‘Original_LC’ (*Mdn* = 4.00), z = -0.49, p = .62, r = .08. All other ratings are significantly different. ‘SRM_Overlay_TRE’ (*Mdn* = 4.00) differs significantly from ‘SRM-overlay_LC’ (*Mdn* = 3.00), z = -2.11, p = .04, r = .36. ‘Motion_TRE’ (*Mdn* = 4.00) received a significantly higher quality rating as ‘Motion_LC’ (*Mdn* = 3.00), z = -2.42, p = .02, r = .62. ‘Label_TRE’ (*Mdn* = 4.00) differs significantly from ‘Label_LC’ (*Mdn* = 4.00), z = -3.67, p = .000, r = .71. ‘Stereo_TRE’ (*Mdn* = 4.00) differs significantly from ‘Stereo_LC’ (*Mdn* = 3.00), z = -4.19, p = .000, r = .41.

8.2.4 Overall Preference

The participants indicated their preference by ranking the different visualization types. This resulted in a preference score ranging from 1 (worst visualization) to 5 (best visualization) (**Figure 12**).

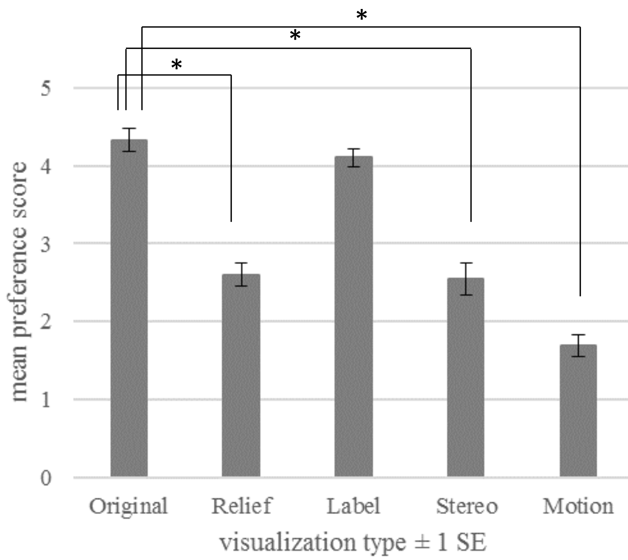


Figure 8.14: Preference between the different visualization types.

A Friedman test showed that the preference of the participants (n = 35) differed significantly between the visualization types ‘Original’ (*M* = 4.33, *SE* = 0.15), ‘SRM-overlay’ (*M* = 2.60, *SE* = 0.15), ‘Label’ (*M* = 4.11, *SE* = 0.12), ‘Stereo’ (*M* = 2.54, *SE* = 0.21) and ‘Motion’ (*M* = 1.69, *SE* = 0.14), $\chi^2(4) = 66.26$, *p* = .000 (**Figure 8.14**). Wilcoxon tests were used to follow up this finding. The preference is significantly higher for the ‘Original’ (*Mdn* = 5.00) than for the ‘SRM-overlay’ (*Mdn* = 3.00), *z* = -4.73, *p* = .000, *r* = .80. Also, the preference differed significantly between the ‘Original’ (*Mdn* = 5.00) and ‘Stereo’ (*Mdn* = 2.00), *z* = -4.05, *p* = .000, *r* = .09. The comparison between ‘Original’ (*Mdn* = 5.00) and ‘Motion’ (*Mdn* = 1.00) is significant as well, *z* = -5.04, *p* = .000, *r* = .68. Only the preference between ‘Original’ (*Mdn* = 5.00) and ‘Label’ (*Mdn* = 4.00) differs not significantly, *z* = -0.52, *p* = .60, *r* = .85. The effect sizes for the significant results are strong whereas the effect size for the non-significant result is small.

8.2.5 Handedness

Accuracy

The participants were classified according to their handedness (right or left). Handedness is expected to be more important with relief maps and therefore for the second study that is not analyzed here. Nevertheless, an exploratory analysis of the handedness in relation to the accuracy of participants with satellite images is conducted here (**Figure 8.15**). Confidence and response time are not expected to be influenced by handedness. Therefore, they are not analyzed here.

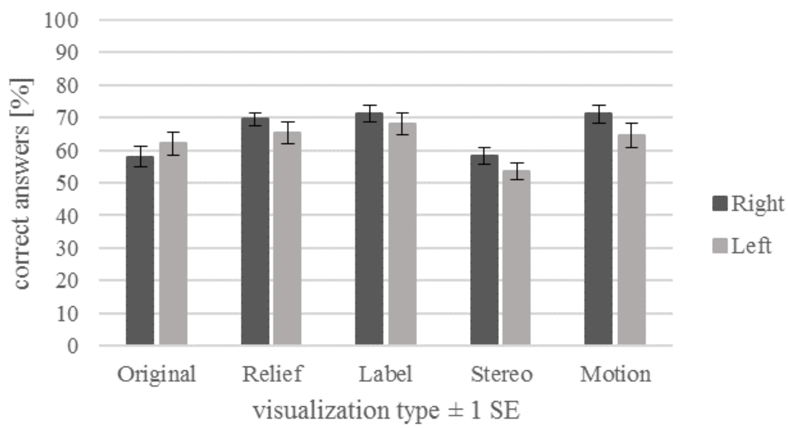


Figure 8.15: Accuracy [%] of different visualization types separated by handedness.

A factorial ANOVA was conducted for the detailed analysis with the factors visualization type and handedness. The result showed that there was no significant main effect of the visualization type on the accuracy of the participant's answers, F (1, 66) = 0.02, p = .90, r =0.02 (**Figure 8.15**). There was a non-significant main effect of handedness on the accuracy, F (1, 66) = 3.27, p = .08, r = 0.22. There was also no significant interaction effect between the visualization type and the handedness of a participant on the accuracy of the answers, F (1, 66) = 2.30, p = .13, r =.0.18.

9 Discussion Experiment II

In this chapter, the results from the previous chapter are analyzed according to research questions 2 and 3 developed in Chapter 2. A summary of the results is given and they are set in relation to current research. Afterwards, limitations are discussed and the main insights are summarized.

9.1 Research Question 2: Correction Methods

Research Question 2*: Empirically, are there differences in participants' accuracy, response time, confidence, quality ratings and preferences with the original satellite image, the best variant of the SRM-overlay correction method from Experiment I as well as with three combinations of the SRM-overlay method (adding labels, stereo respectively motion) for land form and land cover recognition tasks?*

The correction methods are called as follows: SRM-overlay, label, stereo and motion correction.

Accuracy

The results show that there is a significant difference between the original satellite image and the corrections. After a correction is applied, participants answer the TRE-questions on average 75% percent more accurate than before. For the LC-questions the participants answer with the corrected images about 25% less correct than with the original. Ergo, the land cover perception is here less influenced through correction than the land form detection.

The difference of accuracy between task types is large for the original and the stereo correction, but not for the SRM overlay, the label and the motion correction. Generally, a correction should allow the observer to interpret land forms and land cover equally good. Interestingly, for the TRE-questions all corrections are placed on a comparable level between 68% and 73%. In comparison to the original (41%), it can be concluded that all corrections remediate the terrain reversal effect to some extent.

In the **original** image, TRE-questions are answered less accurately than LC-questions. This is expected, as all images were chosen in a way that the terrain reversal effect occurred in all of them. The fact that around 40% answered correctly may be due to experience or learning: some participants might have realized the terrain reversion or some contradictions in the images. Thus, they might have started interpreting the land form with the help of land cover. The amount of accuracy (40.9%) is comparable to the results from Bernabé-Poveda & Çöltekin (2014). They found a mean accuracy of 40.3% for the land form detection in satellite images. For geometrical objects where shading was used as a depth cue, an accuracy of 51% was reported (Baoxia Liu & Todd, 2004). Liu and Todd (2004) also pointed out that the presence of cast shadows and specular highlights would increase the performance. However, this is not applicable to satellite images as the illumination cannot be manually edited after collection. But cast shadows sometimes are available naturally in satellite images when the terrain is highly rugged. According to Liu and Todd (2004), cast shadows should reduce the susceptibility to terrain reversion, but this cannot be confirmed with this study.

The results for **SRM overlay** were similar to the results for label correction because both had the same correction applied. Shading can lead to very different results in accuracy as the reliability and strength relies much on the object viewpoint and changes of the light direction (Lovell et al., 2012; Pentland, 1989). The results depend strongly on the chosen transparency level (Gil et al., 2014; Gil et al., 2010).

Interestingly, the **label** correction performed very well although no additional visual depth cue was integrated. This correction performed equally well as the motion correction and even better than the stereo correction. As motion and stereo are suspected to be very strong depth cues, it can be concluded that also artificial cues have strong effects on our perception. It might be more important how we interpret something and not so much how we perceive it. This again relies back on prior knowledge and biases. As for example Sun & Perona (1998) stated it is not so important where the sun is, but where someone expect it to be.

The findings of Meyer (2015) showed that there was a significant difference in accuracy between **stereo images** and non-stereo images regarding the perception of land forms. These results can be confirmed here. When looking only at the TRE-questions with the stereo correction, answers are significantly more accurate than with the original. Interestingly, the stereo images from this study (*Mdn* = 70.00) led to much more accurate answers than the stereo images in the study of Meyer (2015) (*Mdn* = 32.00). Reasons can be found in two different factors: first, the stereo in this study is combined with the SRM-overlay. As the SRM-overlay corrects the terrain perception and the stereo only adds depth, but does not correct the reversal itself, it can be concluded that the accuracy is much higher because of the SRM-overlay. The second, and probably weaker reason, could be the different task types. Meyer (2015) did not ask LC-questions, but questions that required a more complex interpretation of the image. In contrast to these findings, the stereo image resulted in a poor accuracy rate. The most obvious interpretation for this is that an additional filter was overlaid over the already color reducing SRM. The filter for different colors trigger retinal rivalry (Sexton & Surman, 1999). This, together with the reduced color information from the SRM, reinforces the confounding factors and led to a reduced image interpretation. Interesting findings are also offered for the comparison of SRM overlay, stereo and motion. Although several studies claim that disparity is more reliable than shading for depth perception (e.g. Lovell, Bloj, & Harris, 2012), in the case of terrain correction, stereo could not override shading. A reason for this is mentioned by Lovell et al. (2012): if disparity is influenced by image noise, shading was more reliable. As the use of anaglyph glasses and the SRM overlay might have introduced noise to the image and reduced the reliability of the strong depth cues such as stereo.

According to literature, it was expected that the more cues a scene contains, the better the depth perception is (e.g. Van Beurden, Kuijsters, & IJsselsteijn, 2010). If applied to a terrain reversal correction, this should lead to a better correction than when less cues are available. The SRM overlay contained the shading cue and compared to the motion correction, it can be concluded that motion does not override shading in this case. The same is true for the disparity cue in the **stereo** images. One possibility for these findings is the implementation of these cues. Stereo images were produced with anaglyph images. These are known to introduce difficulties for color perception (e.g. Mehrabi, Peek, Wuensche, & Lutteroth, 2013). However, this fact should only influence the LC-questions. For the TRE-questions a higher accuracy was expected as it was the case in the studies of Meyer (2015). One important difference is that here the satellite image was already corrected with the SRM overlay before stereo was added. If it is considered that both the SRM and the anaglyph view reduced image interpretation quality, it can be concluded that the observer might have experienced a mental overload when looking at stereo images (Westheimer, 2011). Also, the quality of stereoscopic images can be a reason why stereo is not as effective as motion (Van Beurden et al., 2010). Although stereo actually helps perceiving depth (e.g. Meyer, 2015; Van Beurden et al., 2010), in this case too many distractors were present. It seems that the integration of cues does not happen in an additive or multiplicative way in this case. This is similar to what Hubona, Wheeler, Shirah and Brandt (1999) detected. Rather it is possible that complex situations with multiple cues, a linear integration is no longer applied. Previous studies showed that the stereo cue is considered more reliable as the shading cue (Lovell et al., 2012). If this is the case, it is suspected that the confidence in the stereo correction is higher than in the correction where no shading cue is available. However, this cannot be confirmed.

A similar interpretation can be given for the **motion** correction. This correction performs equally well as the SRM overlay although it is claimed that with motion a powerful depth cue can be integrated (Vezzani et al., 2015). The depth cue integration probably was hindered in the motion correction due to several implementation facts. First, the motion was implemented as object motion. Other studies found that object motion was equally suited to perceive depth as motion parallax (Van Beurden et al., 2010; Willett et al., 2015). However, with the object motion the mental workload can be smaller and the visual comfort higher than with motion parallax (Van Beurden et al., 2010). However, it plays a major role if motion is introduced interactively or as an animation. Interactive displays are known to perform better, but also introduce new challenges (Willett et al., 2015). Willett et al. (2015) say that 3D visualizations need much more complex interactions and therefore they are more difficult to interpret for humans. In contrast to this, humans tend to prefer simpler representations over more complex ones (Brügger, 2015; Willett et al., 2015). Here, we displayed an animation of an object motion and therefore, the motion cue probably was not optimally displayed. Another major drawback is that animation cannot be used in most situations outside the laboratory (Willett et al., 2015). However, another implementation would increase the effort in a probably exponential way and it is not assured that the cost-benefit-ratio is small. Additionally, motion was experienced both amusing and disturbing. This can be concluded from the reactions of the participants during the study. An increase of accuracy might be generated when combining stereo an object motion (Van Beurden et al., 2010). Interestingly, motion could not outperform the stereo correction indicating that in this case, these cues were similarly strong as it is also found by Liu and Todd (2004), Řeřábek et al. (2011) and Todd and Norman (2003).

The accuracy results for label and motion correction were similar to the results of Bernabé Poveda & Çöltekin (2014) for 180° rotated images (72%). However, label and motion keep the north orientation.

In comparison to Experiment I, the accuracy for the original in Experiment II was lower for TRE-questions (-10%) and higher for LC-questions (+8%). This can rely on the fact that in Experiment II only non-experts were chosen to participate. For the SRM-overlay with 65% transparency, an inversed tendency was observed (+3% for TRE, -6%), but to a smaller extent. It can therefore be concluded, that expertise might affect the perception of the terrain reversal positively, but not the land cover. This effect can be mitigated with the SRM-overlay correction.

It can be concluded from the accuracy analysis that the SRM-overlay correction extracted from Experiment I could not be enhanced using additional depth information. Rather, land cover was impoverished in perception as for example in the case of adding stereopsis. Nevertheless, all corrections performed significantly better than the original. The SRM-overlay, label and the motion corrections can be accepted as valid corrections although the overall accuracy never reached 80% or more. It is important to know that results of accuracy for different visualization depend on the choice of stimuli and design and therefore a comparison is not always possible (Lovell et al., 2012). Also, many studies investigated geometric forms and did not use satellite images and therefore, the results are not directly comparable.

Response Time

There is no significant difference in response time between the original and any correction type, even if both task types are reported separately. These results are in line with the findings of Meyer (2015). An explanation for this can be found in the instructions that were given to the participants. Similar as with Meyer (2015), the participants were told to solve the tasks quickly in order to extract what the participants see and not giving them time to interpret the images. However, no time limit was set. Although there is no significant result, participants had longer to answer the questions for the LC-questions with any corrections. This again can be explained by the reduced color perception. Although the results were not significant, the next section discusses some of the nuanced differences that were found.

Generally, participants are faster in answering TRE-questions with the SRM overlay and the motion correction than with the original. In contrast, they are slower when using label or stereo correction. For the LC-questions, participants are only faster with the SRM overlay, but slower with the other correction methods. As the completion time depends a lot on the difficulty of the task (Van Beurden et al., 2010), it was tried to balance difficult and easy images.

The participants needed the most time for the LC-questions with the stereo correction. This corresponds to the accuracy-response time trade-off where individuals answer more correctly but also need more time to solve a task (Hubona et al., 1999). This is in line with the findings of Van Beurden et al. (2010). However, they also found no significant change in accuracy with adding stereoscopy. Another reason for the larger response time might be the fact that the participants had to handle two overlays: the SRM and the filter for the anaglyph images. Also, with anaglyph glasses the eyes need some time to adapt themselves for perceiving depth in this way. This suggestion can indirectly be confirmed with the findings of Meyer (2015): she found no significant difference between response times for stereo and non-stereo images. However, the mean response time for stereo images was little shorter than for non-stereo images (Meyer, 2015). Another reason for this finding might be the instruction to solve the task as quickly as possible. For the label correction, the slow responses can be explained because the participants here started to interpret the task or were distracted from the fact that not only pictorial information was available. It is suspected that written text is processed differently than pictorial information which can lead to a larger response time. For the SRM overlay participants were faster for both task types. For the TRE-questions it can be concluded that they perceived depth better, but they were neither overwhelmed with additional information nor with mental challenging images. Interestingly, participants answered the TRE-questions faster with the motion correction than with the original. This might be because motion actually adds some depth information. This is possible as the accuracy analysis shows that participants answered more correctly with the motion correction. These results are in contrast to what Van Beurden et al. (2010) found. They state that object motion increases the completion time significantly. However, it could also be that the animated images appeared annoying or disturbing and therefore, the participant wanted to rush through this section. It could also be that the motion was disturbing and the participant started guessing the answer. However, this statement can neither be accepted nor rejected.

Overall, the response time can also be influenced by the learning effect that inevitably occurred despite randomizing the order of the images. Therefore, it might be that some participants answered questions according to what they remembered from previous images. Also, the standard error was rather large for all visualizations indicating some variability in response time among the participants. This can derive from individual differences. It can be concluded that the SRM overlay and the motion correction were answered little, but not significantly, faster than the rest of the visualizations.

Confidence

There is no significant difference found in the confidence of the participants between visualization types. These results are in line with what Meyer (2015) found. However, the difference was significant within the visualization types and between the task types. Generally, participants were significantly more confident with the LC-questions than with the TRE-questions. A reason for this might be the different question types and the artificial rating system. For TRE-questions it is more likely that participants are classified as 'unsure' because on the Likert scale only the levels 1 and 5 are rated as completely confident. In contrast to this, at the LC-questions there was only one possibility to be rated as not confident – when the participant marked 'not sure / ambiguous' as answer. Also, participants might be more familiar with classifying landscapes into thematic classes than with judging land forms. While land cover can also be estimated, luckily guessed or interpreted according to the surrounding, it is harder to do this with a land form where there is more or less a 50% chance of getting the correct answer.

If corrections are aggregated, there is a significant drop in confidence for LC-questions in comparison to the original but not for TRE-questions. If corrections are split up, it becomes visible that only with the motion correction participants were less confident compared to the original. This drop in confidence obviously influences the aggregated correction. Interestingly, motion stands out again in the case of confidence as this is the only visualization that has no significant difference within the correction method itself. This means that there is no large difference in confidence between TRE-questions and LC-questions with the motion correction.

From the confidence analysis it can be concluded that participants answered comparably confident throughout the different visualization. This indicates that the participants were not aware of the terrain reversal effect during the completion of the study. This suggestion stands in contrast to the indication of the participants: 49% of the participants stated after the study that they realized the terrain reversal effect. It might be that participants only realized afterwards with what they were confronted but they still indicated that they realized the terrain reversal effect. Another reason for this discrepancy is the calculation of confidence with the Likert scale. The only correction method that leads to a significant loss of confidence is the motion correction in the case of LC-questions. This method therefore falls behind the others regarding confidence.

Quality

The participants rated the quality of the different visualizations significantly different. The participants rated the original generally with higher quality than the corrections, but only the difference between the LC-questions in the original and the LC-questions in the corrections was significant. The mean rate for the original image lies over 4 out of 5 and is therefore a rather high rating (in words: “good”). As expected, the participants rated the quality to answer LC-questions much worse when using corrections and on average they gave a score little higher than 3 out of 5 (in words: “acceptable”). When looking at the single visualizations, it can be concluded that the original was the best rated visualization in terms of quality. This was also the case in Experiment I. A reason for this might be the familiarity for the participants as this is the visualization they have most frequently seen until now (Brügger et al., 2016). Also, the mental overload for these images is clearly not as high as with the corrections. Within the corrections, all LC-questions were rated with lower quality than the TRE-questions. The SRM-overlay and the motion correction can be put in one group: TRE- and LC-questions were similarly rated, but both lower than the original. For the label and the stereo correction, discrepancy between TRE- and LC-questions is higher, with a TRE-rating that is only little worse than for the original. Stereo clearly hindered a good color classification due to the overlay of more than one layer. Also, stereo images are suspected to lead to discomfort according to Mehrabi et al. (2013). This might have an influence here, although the majority of the participants indicated that they did not feel any or only little discomfort with stereo images. Motion obviously disturbs the participants less. In summary, the original images are rated best, but all corrections except for the land cover experience with stereo images are rated at least acceptable. This indicates that the participants do not rate visualizations very well although they perform more accurately with them.

Preference

The preference for a visualization type by the participants corresponds only partially with the quality rating. They also liked the original the most, followed by the label correction. SRM overlay and stereo correction yield the same preference score and motion was the least preferred visualization. The latter is surprising, because the quality rating showed different results. It might be that the shivering animation was annoying to look at for a longer time period and therefore the participants did not like this visualization type. However, the amount of discomfort should not be overestimated: in other studies, object motion was found to cause less discomfort than other visualizations (Van Beurden et al., 2010) and in this study, participants did not indicate any discomfort. The preference result do not correspond to the findings of

Hegarty et al. (2009). While our result suggest that the least adapted images are preferred, they state that individuals tend to prefer enhanced visualizations that included more complex visual task such as animations or realism (Hegarty et al., 2009). Realism is also a part of satellite images, but nevertheless, they were the favored visualization. A reason for the findings that motion and stereo were not preferred might be that they were hard to perceive. The animations shown in our studies were complex as they were combined with a layer that additionally lowered the surface information. As no interactivity was included, it might be that the comprehension of the scene was severely hindered (Hegarty et al., 2009). This argument is strengthened by our results for confidence: the motion correction was the only visualization where participants answered significantly less confident. It can be concluded that participants like what they are already used to and do not like what causes more mental working process. If they have the possibility to select from several alternative visualizations, it is likely that they choose what is easy and reliable and not a display that is requires more effort to understand (Hegarty et al., 2009).

Research Question 2*: Empirically, are there differences in participants' accuracy, response time, confidence, quality ratings and preferences with the original satellite image, the best variant of the SRM-overlay correction method from Experiment I as well as with three combinations of the SRM-overlay method (adding labels, stereo respectively motion) for land form and land cover recognition tasks?*

In summary, the research questions can be answered as follows: There are several significant differences in participants' accuracy and few significant effects influencing their confidence. The response time is not influenced, but quality and preference ratings differed significantly. The accuracy of the participants' answers proceeds reversed for land form and land cover detection. According to quality and preference findings, participants prefer visualizations where they are confident and which are suited for a correct land cover detection, but not for land form recognition. The second hypothesis stated that participants perform better in land form recognition tasks with combined methods than with simple ones. The hypothesis is rejected and reformulated: participants performed better with combined correction methods in comparison to no correction at all (i.e. the original image). However, the combined correction methods did not let the participants perform significantly better than the simpler SRM-overlay. The third hypothesis suggested that participants perform equally well with all correction methods, but better with the original satellite image in terms of land cover tasks. This hypothesis can be partially accepted. The original image performed in fact better than the corrections and was also preferred as visualization type. However, corrections were not equally among each other. Motion led to a reduced confidence level and both stereo and motion yielded lower accuracy rates and were not preferred from the participants.

9.2 Research Question 3: Handedness

***Research Question 3**: Is there a difference between left-handed and right-handed participants regarding their overall accuracy with the different correction methods mentioned in Research Question 2?*

Accuracy

Handedness was taken into account without a specific expectation of the results. However, from literature it can be suggested that handedness can have an influence on the accuracy of the participants (Sun & Perona, 1998). No influence is expected for confidence, response time, quality or preference and therefore, these factors are not analyzed. The findings showed that there was no significant difference in accuracy between right- and left-handed participants. This is in line with the findings of Andrews et al. (2013). Interestingly, the left-handed participants answered little more accurate with the original, but with all correction types, the right-handed participants answered more correctly. However, these results probably should not be related automatically to the handedness of the participants. It is highly possible that other factors influenced this result. For example, more women were in the left-handed group than in the right-handed. Also, the spatial ability was higher in the right-handed group which can be related to the fact that more men were in the right-handed group than women (e.g. Lawton, 2009). It is not clear if and how the results are influenced by any of these factors and it is possible that the discrepancy is arbitrary. Possible confounding factors are not further investigated here and it can be concluded that there is no difference in accuracy between right- and left-handed participants. There is one important distinction between the findings of Andrews et al. (2013) and Sun and Perona (1998) and the results from this experiment: the other studies used artificially lit geometric figures whereas this experiment here uses satellite images that are overlaid with a SRM. This is possibly the reason that no significant differences in accuracy were found between right- and left-handed participants.

In summary, the third research question can be negated as there is no difference in the accuracy of participants. Consequently, it is hypothesized that handedness has no influence on the accuracy of individuals with land form or land cover tasks when using satellite images that are combined with a corresponding semi-transparent SRM as well as other depth information.

9.3 Other Findings

Spatial Ability

The spatial ability of an individual can be related to the extraction of three-dimensional structure out of two-dimensional images (Lobben, 2007). It might be relevant for perceiving depth in two-dimensional satellite images and is therefore investigate here. The accuracy in the spatial ability test correlated with the SRM overlay correction (both task types aggregated), the label correction (also aggregated task types) and with the response time for the original (also aggregated task types). The correlation of the spatial ability with the accuracies is positive, while the correlation with the response time is negative. This can give indications that a higher spatial ability helps with interpretation of shading as depth cue. The same applies to artificial cues, but not to other pictorial depth cues. This means that with higher complexity in the images, the participants with higher spatial abilities appear to no longer benefit from their advantage. The results for the response time suggest that participants with higher spatial abilities were faster with the original satellite image. But again, they could not rely on this advantage with the corrected images. Maybe more complex images introduce more distraction and familiarity is reduced. Brügger (2015) investigated the spatial ability of participants for path choices and found no significant effect. Therefore, and due to the above mentioned reasons, caution is appropriate when interpreting the results. Most participants indicated that they experienced the test as very difficult. They implicitly said that some of their answers were based on guessing. Additionally, they had a time limit of six minutes. This might have added another source of stress despite the bad feeling they had when confronted with a task they found enormously difficult. It is therefore possible that the results were distorted.

Stereoscopic Vision

It was investigated if stereoscopic acuity was related with the performance of the participants when they used the stereo correction. In this case, the land form detection is important as stereopsis is expected to enhance depth perception. However, no correlation is found between the amount of correct answers for TRE-questions with stereo correction and the stereo acuity of the participants. The preceding test shows that all participants are able to see stereoscopically. It can therefore be concluded, that the amount of stereoscopic vision might not be as important for the correction with stereo as the factor to have stereo vision at all. However, it might be more important when other methods instead of the anaglyph stereoscopy would have been used. In comparison to Hess, To, Zhou, Wang, & Cooperstock (2015), the results correspond to what they call “the haves”, i.e. the group of participants that have good stereoscopic vision.

9.4 Limitations

In summary, the studies show that correction methods differ significantly from the original satellite image. However, the interpretation is complicated due to the fact that good conditions for the correct land cover detection seem to disagree with what would be good to discover land cover classes. The results could also be influence by several limitations such as the study design or the implementation. These limitations are discussed in the following section.

Participants

In the online study, 93 participants took part (Experiment I) and 35 participants completed the controlled lab study (Experiment II). Although this is in both cases a sufficiently large number to perform statistics (Field, 2009), more participants might have led to different results. Especially the results for accuracy and response time might differ when more participants take part. Also, when participants were classified in two parts, the left-handed group only contained 15 persons. It also would have been good if both group contained an equal number of participants as this would be easier to compare the two groups statistically (Field, 2009). In Experiment II only non-experts, i.e. people without professional geographic education or background, participated. This does not allow a comparison according to expertise. Additionally, a lab environment does not represent the environment were satellite images are usually used. Although they are used frequently on a computer, they could also be used in the form of hard copies or on mobile devices. The different displays were not considered in this experiment.

Learning Effect

The different visualization types were randomized to minimize a possible learning effect. However, as only 10 distinct satellite images were chosen, each participant saw the same location 10 times (5 visualization types, 2 task types). Many participants also mentioned in the end of the study that they recognized some images and that this might have changed their answer. However, these qualitative estimates were neither assessed systematically nor deeper discussed. The multiple repetitions of the images might have led to tiredness or cumbersomeness. This potentially influences the attention of the participant and therefore has an impact on the given answers. However, as the randomization worked properly (in contrast to Meyer, 2015), the influence of tiredness should be balanced.

Visualizations

The stimuli were chosen according to subjective criteria. As the perception of the terrain reversal effect varies between individuals (e.g. Bernabé Poveda & Çöltekin, 2014), it is possible that some participants did not see any effect in an image. It is therefore difficult to estimate if and to which extent the findings can be generalized as other visualizations might have led to different results. Another limitation of the stimuli is the implementation, especially for the SRM overlay, the stereo and the motion correction. For the SRM overlay, a color correction similar to how Bernabé-Poveda, Sánchez-Ortega, & Çöltekin (2011) suggested would have probably led to results with higher quality for interpretation. The stereo images were created for the use of anaglyph glasses which impeded the visualization interpretation furthermore (see section *Anaglyph Glasses*). The motion correction was presented in a form of animated object motion. This introduced an annoying shivering effect which might have distracted the participant from the actual task. An interactive image as it is presented with the elastic terrain (Jenny & Buddeberg, 2016) might have added more useful depth perception and affected the performance of the participants more. Another method which is similar to the elastic terrain is the wiggle option presented by Řeřábek et al. (2011). They show that this technique is clearly preferred over anaglyph stereo and their results show that it is also much more preferred than the object motion animation. Due to technical and time-wise limitations, this was not implemented here.

In addition, artificial marks were added to the satellite images in order to solve the different tasks. They might have changed the perception of the participants in terms of color interpretation or land form detection. They might have also covered features from the image. However, they were chosen in a way that they did not cover important information. It is difficult to estimate the influence of these factors based on the answers of the participants.

Anaglyph Glasses

A major disadvantage of anaglyph glasses is that individuals loose color information (Mehrabi et al., 2013). This affects the interpretation of the image severely. In these studies, the land form detection and the land cover detection were weighted as equally important and therefore the interpretation of colors was important for the overall evaluation of the correction method.

Task types

Another restriction was the scale of the chosen stimuli. The satellite images were zoomed to one specific land form which automatically reduced the amount of information. However, in an everyday situation where satellite images are used, participants probably see more than only one land form. This means that the tasks were actually more difficult than in a situation outside of the laboratory. Both task types were equally weighted that means land cover perception was as important as land form detection. If this is also the case in non-laboratory conditions is questionable.

Accuracy, Confidence, Quality and Preference

The accuracy of the participant's answers might have been influenced by the global convexity bias. It describes the prior to perceive land forms rather as convex than concave (Langer & Bülthoff, 2001; Baoxia Liu & Todd, 2004). Therefore, participants may judge land forms rather as hills especially when the terrain is ambiguous. The influence of this effect on accuracy was also discussed by Liu & Todd (2004) and found from Aubin & Arguin (2014). The confidence was indirectly measured through the Likert scale. Asking the participants directly after each question how confident they felt, might have led to a different confidence distribution. However, also asking repeatedly about confidence would influence the confidence of the participant and might lead to uncertainties. The preference of the participants was only asked for aggregated visualization types, but not separated after task type. This limits the amount of interpretation possibilities. Also, more information might be drawn from asking participants twice about their preferred visualization, i.e. before and after the study is conducted (Brügger et al., 2016). Additionally, preference might be linked to spatial ability (Brügger et al., 2016; Hegarty et al., 2009). This interaction was not evaluated in this study.

Instructions

The participants were instructed to solve the tasks as fast as possible but without a hurry. The aim was to receive answers about what they perceived and to not let them too much time to interpret the images. However, a separation between perception and interpretation is hardly possible. Especially for the label correction, participants were in a conflict between what they interpreted and the request to answer questions according to what they see. The time limitation might have also influenced the accuracy and confidence of the participants. Although they did not have a time limit, they might have been put into unconscious stress by telling them to answer as quickly as possible.

Expertise

For Experiment I, the experience level of the participants was explored in order to extract information about their expertise. However, participants were not directly asked about their profession and the expertise level was directly derived from the experience indications. This definition might not be as accurate as when participants were asked directly. Non-experts were selected for the second experiment because they represent a majority of map users. A limitation of this choice is that no comparisons to experts were made.

Usability

The usability of the correction methods is an important factor in judging if a correction method is suitable. This is preceded by the feasibility of the method. The SRM-overlay can only be applied when a DEM is available. The labelling correction depends a lot on the language and if the label includes a land cover cue (e.g. 'river', 'road' etc.). Stereo and motion corrections require more effort to create stimuli and they are also only usable with restrictions. Stereo images require stereoscopic vision and are not easy to implement depending on the mode (Mehrabi et al., 2013; Sexton & Surman, 1999). Correcting with motion cannot be used with hard copies, but as mobile devices are wide spread, these provide a feasible alternative. In summary, comparing to the original, some correction methods include larger effort than others which influences their usability.

9.5 Summary

The experience of quality might be related to confidence and accuracy as it can be suggested that if participants rate a visualization with good quality, they must have thought that they can answer accurately and they were also confident about it. For the original visualizations of the TRE-questions the rating of quality is high as is the confidence, but accuracy is low. This indicated that participants were not aware of the fact that they answered wrongly. With the LC-questions for the original, quality is high and confidence is even higher as well as the accuracy. Here it is not the case that participants rate better than they performed. In contrast, when using corrections, the quality for the TRE-questions is rated lower, but the confidence and especially the accuracy are higher. For the LC-questions in the correction, quality is low, confidence is high and accuracy is also lower than with the original.

In **Table 9.1** an overview of the discussed correction methods in comparison to each other and to the original is given. It can be concluded that regarding land form detection, subjective experience often stood in contrast to the performance. This is not the case for land cover perception because LC-questions were often answered accurately. Performance and subjective experience of the participants are on an equally high level. This emphasized the unawareness of the terrain reversal effect in our visual processing system and relates back to the prior that guide our brain's interpretations. The thesis is led by the question 'Which correction method remediates the terrain reversal effect in satellite images and also preserves the interpretation of them best regarding the performance and subjective experience of participants?'. Due to the multiple strengths and weaknesses of every method, it is not possible to choose one single method that fits most conditions. While generating images, there is a constant search for the best compromise between good perception of form and communicating other information as for example about the land cover (Willett et al., 2015). The choice for the best fitting correction method is dependent on the purpose of the visualization and the tasks that the individuals want or have to accomplish. Thereby, image providers as well as the users have to be aware that perceptual issues exist and that any influences on the familiar appearance might influence the performance and preference of the users. In terms of the terrain reversal effect, further research is needed to investigate well-balanced methods that are both effective and applicable.

Table 9.1: Summary of the evaluated correction methods. + means significantly better than the original, - stands for significantly worse. 0 means no significant difference was found between a correction method and the original.

Original	*Task Type*	**Accuracy**	**Response Time**	**Confidence**	**Quality**	**Preference**
SRM-Overlay	*TRE*	+	0	0	0	-
	LC	-	0	0	-	
SRM-Overlay + Label	*TRE*	+	0	0	0	0
	LC	-	0	0	-	
SRM-Overlay + Stereo	*TRE*	+	0	0	0	-
	LC	-	0	0	-	
SRM-overlay + Motion	*TRE*	+	0	0	0	-
	LC	-	0	-	-	

10 Conclusion

This thesis investigated several correction methods for the terrain reversal effect in satellite images. The aim of the project was to gain more knowledge about this specific perceptual phenomenon in the context of depth extraction from 2D images. To achieve this goal, a selected set of promising correction methods was applied to different satellite images. The method that achieved the best results regarding accuracy, response time, confidence, quality ratings and preference for selected image interpretation tasks was considered to be the most promising correction method.

Terrain Reversal Correction

The results report a partially higher performance of the participants with satellite images on which a correction method was applied than with original images. More specific, the accuracy of the participants' answers is higher with all applied correction methods, but the response time does not differ compared to original image. These results confirm that illumination is a crucial factor for identifying spatial relationships and land forms in satellite images. The interpretation of satellite images is guided by the assumption of an overhead light source (Cavanagh & Leclerc, 1989; Kleffner & Ramachandran, 1992; Saraf et al., 2007). When applying the SRM-overlay method, attention is payed to this assumption and the light source is shifted towards a position that allows a top-left illumination (Gil et al., 2014; Saraf et al., 1996). Interestingly, a raise of the number of depth cues does neither result in a higher accuracy nor in a faster completion of tasks. These findings support the theory that depth cues are not integrated linearly when they co-occur in an image (Bülthoff & Mallot, 1988; Landy et al., 1995; Vuong et al., 2006). An additive or multiplicative approach seems insufficient to explain how the visual system processes multiple depth cues. Within the scope of this theory, it can be observed that single depth cues veto each other depending on their relative reliability in each satellite image scenery. The findings rather provide evidence that in complex images, depth perception relies on the integration of cue in a non-linear way.

Influence on Land Cover Interpretation

The different correction methods to counter the terrain reversal effect influence the outcome of land cover recognition tasks. This corresponds to the major concern of several previous studies is that the SRM-overlay method reduces color and texture information (Bernabé-Poveda et al., 2005; Gil et al., 2014; Saraf et al., 2005). The results of the studies from this project have shown that participants answered land cover questions less accurately with visualizations that contained any of the correction methods mentioned above. Interestingly, the reduced performance for land cover detection does not influence the confidence level of the participants. They were as confident with their answers as with the original images though they performed more accurately with the latter in land cover tasks. Furthermore, participants did not need more time to answer questions with visualizations where the information extraction was more challenging due to one or more overlays. Both findings are confirmed with the quality ratings and preference levels: the unimpeded visualizations are favored.

Performance vs. Subjective Experience

The results for land form as well as land cover detection suggest that adding more depth information leads to more complex visualizations. This, together with the finding that confidence cannot be increased with any of the correction methods, indicates that adding complexity might reduce the performance and subjective experience of participants. These findings are strengthened with the analysis of preference and quality: participants tend to prefer more realistic looking images. Hence, if individuals can choose different alternatives, they would probably pick the most familiar and easily interpretable one. This choice is not necessarily corresponding to the visualization with which individuals perform best (Brügger et al., 2016; Hegarty et al., 2009).

Recommendations and Future Research

A successful interpretation of satellite images involves the correct recognition of land forms and land cover. Only if both parts are equally well interpretable, the visualization is useful for a large range of people. Although the ability to perceive land form and land cover can be investigated separately for each visualization type, these two tasks cannot be split up in actual applications of satellite images. The analysis of the results indicates that these two tasks sometimes contradict each other in terms of the participants' performance and subjective experience. It is therefore a requirement that an application-oriented method to correct the terrain reversal effect integrates both abilities equally well. Besides familiarity with visualizations, preference and performance, other elements influence the application of correction methods. Individual factors, various image sceneries and different depth cue combinations make it impossible to select one single correction method with the aim to correct the terrain reversal effect in all possible situations. Nevertheless, the results of the two studies in this thesis allow to make recommendations for the suitability of the examined correction methods:

- If the advantages and disadvantages of the different method investigated in this project are cumulated and analyzed in absolute terms (see section 9.5, Table 1), none of the methods is recommended to apply one-to-one in satellite images in order to correct the terrain reversal effect.

- The SRM-overlay method as a possible approach to correct the terrain reversal effect is still a useful method to correct *land forms*.
- Adding labels increases the preference of the participants. It is therefore recommended to use additional artificial information in visualizations. The optimal amount of information is subject to further research.
- The use of stereopsis with anaglyph glasses in order to enhance depth is not recommended as this method introduces several disadvantages and only presents a reduced applicability. Another implementation of stereoscopic displays might lead to better results.
- The application of motion for the correction of the terrain reversal effect cannot be recommended according to the findings of this project. However, this method is not yet entirely investigated and especially in the field of terrain reversion, more research is required in order to assess the actual potential of motion as a helping tool for correct land form detection.

Based on the analysis of the results, it can be concluded that current correction methods need further investigations in order to reduce their disadvantages and optimize their useful features. One key issue with currently proposed methods is the harmful reduction of color information after the application of the correction. Further research should focus on technical possibilities to maintain the spectral information of the satellite image without the loss of classification or interpretation options. Technical improvements could also lead to enhanced implementations of motion. Especially, the use of mobile devices instead of fix installed computers or hard copy maps offer new possibilities to show maps and satellite images (e.g. Řeřábek et al., 2011). These alternative display modes need to be considered when correcting the terrain reversal effect because it has an influence on the usability of the method. Further research could address these options. Furthermore, understanding the interaction between performance and subjective experience of individuals is important for evaluating correction methods. This might be even more interesting to investigate when interactivity is added to the visualization display (Řeřábek et al., 2011; Willett et al., 2015). Finally, two application-oriented issues could be subject to further studies: on the one hand, the feasibility of a developed method should be considered. Not only is it important that a method actually works, but also that it stays in balance with costs and benefits. On the other hand, a key to a promising method is its usability for as many individuals as possible especially outside a laboratory environment.

Bibliography

ADA (2007). Dental admission testing program – Sample Test Items. American Dental Association.

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the “light-from-above” prior. *Nature Neuroscience*, *7*(10), 1057–8. http://doi.org/10.1038/nn1312

Admission, D., & Program, T. (2007). Dental Admission Testing. *Time*, 6–7.

Anderson, J. R., Hardy, E. E., Roach, J. T., Witmer, R. E., & Peck, D. L. (1976). A Land Use And Land Cover Classification System For Use With Remote Sensor Data. *A Revision of the Land Use Classification System as Presented in U.S. Geological Survey Circular 671*, *964*, 41.

Andrews, B., Aisenberg, D., D’Avossa, G., & Sapir, A. (2013). Cross-cultural effects on the assumed light source direction: evidence from English and Hebrew readers. *Journal of Vision*, *13*(13), 2. http://doi.org/10.1167/13.13.2

Aubin, M., & Arguin, M. (2014). Stereo and shading contribute independently to shape convexity–concavity discrimination. *Perception*, *43*(4), 333–343. http://doi.org/10.1068/p7568

Bernabé-Poveda, M. a., Sánchez-Ortega, I., & Çöltekin, A. (2011a). Techniques for Highlighting Relief on Orthoimagery. *Procedia - Social and Behavioral Sciences*, *21*, 346–352. http://doi.org/10.1016/j.sbspro.2011.07.028

Bernabé-Poveda, M. a., Sánchez-Ortega, I., & Çöltekin, A. (2011b). Techniques for Highlighting Relief on Orthoimaginery. *Procedia - Social and Behavioral Sciences*, *21*, 346–352. http://doi.org/10.1016/j.sbspro.2011.07.028

Bernabé-Poveda, M., Callejo, M., & Ballari, D. (2005a). Correction of Relief Inversion in Images Served by a Web Map Server. In *International Cartograhic Conference (ICC)* (Vol. 1). Retrieved from http://www.cartesia.org/geodoc/icc2005/pdf/oral/TEMA6/Session 8/ANGEL BERNABE.pdf

Bernabé-Poveda, M., & Çöltekin, A. (2014). Prevalence of the terrain reversal effect in satellite imagery. *International Journal of Digital Earth*, *0*(0), 1–24. http://doi.org/10.1080/17538947.2014.942714

Bernabé-Poveda, Callejo, & Ballari, D. (2005b). Correction of relief inversion in images served by a web map server. In *International Cartograhic Conference (ICC)* (Vol. 1). Retrieved from http://www.cartesia.org/geodoc/icc2005/pdf/oral/TEMA6/Session 8/ANGEL BERNABE.pdf

Bharathi, R. D., & Priyadharshini, C. I. (2016). ENHANCED VISION OF HAZY IMAGES USING IMPROVED DEPTH ESTIMATION AND COLOR ANALYSIS, *3*(3), 249–253.

Biland, J. (2014). Terrain reversal effect : Assessing the impact of incidence of light in shaded relief maps and comparison of landform perception in shaded relief maps and satellite images.

Biland, J., & Çöltekin, A. (2016). An empirical assessment of the impact of the light direction on the relief inversion effect in shaded relief maps : NNW is better than NW, *406*(June), 1–24. http://doi.org/10.1080/15230406.2016.1185647

Braunstein, M. L. (1962). The perception of depth through motion. *Psychological Bulletin*, *59*(5), 422–433. http://doi.org/10.1037/h0044239

Brügger, A. (2015). WHERE ARE THE UPS AND DOWNS ? Evaluating Elevation Representations for Bicycle Paths in City Maps.

Brügger, A., Fabrikant, S. I., & Çöltekin, A. (2016). An empirical evaluation of three elevation change symbolization methods along routes in bicycle maps. *Cartography and Geographic Information Science*, *406*(June), 1–16. http://doi.org/10.1080/15230406.2016.1193766

Bülthoff, H. H., & Mallot, H. a. (1988). Integration of depth modules: stereo and shading. *Journal of the Optical Society of America. A, Optics and Image Science*, *5*(10), 1749–58. http://doi.org/10.1364/JOSAA.5.001749

Bülthoff, I., Bülthoff, H., & Sinha, P. (1998). Top-down influences on stereoscopic depth-perception. *Nature Neuroscience*, *1*(3), 254–7. http://doi.org/10.1038/699

Cavanagh, P., & Leclerc, Y. G. (1989). Shape from shadows. *Journal of Experimental Psychology. Human Perception and Performance*, *15*(1), 3–27. http://doi.org/http://dx.doi.org/10.1037/0096-1523.15.1.3

Çöltekin, A., Fabrikant, S. I., & Lacayo, M. (2010). Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings. *International Journal of Geographical Information Science*, *24*(10), 1559–1575. http://doi.org/10.1080/13658816.2010.511718

Çöltekin, A., Heil, B., Garlandini, S., & Fabrikant, S. I. (2009). Evaluating the Effectiveness of Interactive Map Interface Designs: A Case Study Integrating Usability Metrics with Eye-Movement Analysis. *Cartography and Geographic Information Science*, *36*(1), 5–17. http://doi.org/10.1559/152304009787340197

Çöltekin, A., Lokka, I., & Boér, A. (2015). The Utilization of Publicly Available Map Types by Non-Experts – A Choice Experiment.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. http://doi.org/10.1038/415429a

Field, A. (2009). *Discoveing Statistics using SPSS* (Vol. 25). http://doi.org/10.1007/s13398-014-0173-7.2

Gargantini, A., Facoetti, G., & Vitali, A. (2014). A Random Dot Stereoacuity Test based on 3D Technology. *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, *5*, 3–7. http://doi.org/10.4108/icst.pervasivehealth.2014.255333

Gerardin, P., de Montalembert, M., & Mamassian, P. (2007). Shape from shading: New perspectives from the Polo Mint stimulus. *Journal of Vision*, *7*(11), 1–11. http://doi.org/10.1167/7.11.13

Gerardin, P., Kourtzi, Z., & Mamassian, P. (2010). Prior knowledge of illumination for 3D perception in the human brain. *Proceedings of the National Academy of Sciences*, *107*(37), 16309–16314. http://doi.org/10.1073/pnas.1006285107

Gil, M. L., Arza, M., Ortiz, J., & Avila, A. (2014). DEM shading method for the correction of pseudoscopic effect on multi-platform satellite imagery. *GIScience & Remote Sensing*, *51*(6), 630–643. http://doi.org/10.1080/15481603.2014.988433

Gil, M., Ortiz, J., Rego, T., & Gelpi, L. (2010). The Correction of the Pseudoscopic Effect on Quickbird Satellite Imagery. *Survey Review*, *42*(318), 318–326. http://doi.org/10.1179/003962610X12747001420465

Goldstein, E. B. (2002). Perceiving Depth and Size. *Sensation and Perception*, (1), 225–268.

Häkkinen, J., & Gröhn, L. (2016). Turning water into rock: The inverted waves effect. *I-Perception*, *7*(1). http://doi.org/10.1177/2041669515627951

Hartle, B., & Wilcox, L. M. (2016). Depth magnitude from stereopsis: Assessment techniques and the role of experience. *Vision Research*, *125*, 64–75. http://doi.org/10.1016/j.visres.2016.05.006

Hegarty, M., Smallman, H. S., Stull, A. T., & Canham, M. S. (2009). Naïve Cartography: How Intuitions about Display Configuration Can Hurt Performance. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *44*(3), 171–186. http://doi.org/10.3138/carto.44.3.171

Hess, R. F., To, L., Zhou, J., Wang, G., & Cooperstock, J. R. (2015). Stereo vision: The haves and have-nots. *I-Perception*, *6*(3). http://doi.org/10.1177/2041669515593028

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of Some Resistant Performance Rules for Labeling Outlier. *Journal of the American Statistical Association*, *81*(396), 991–999. http://doi.org/10.1080/01621459.1986.10478363

Horn, B. (1989). Obtaining Shape from Shading Information. *Shape from Shading*, 123–171.

Hubona, G., Wheeler, P. N., Shirah, G. W., & Brandt, M. (1999). The Relative Contributions of Stereo , Lighting and Background Scenes in Promoting 3D Depth Visualization 1, 1–25.

Humphrey, G. K., Goodale, M. A., Bowen, C. V, Gati, J. S., Vilis, T., Rutt, B. K., & Menon, R. S. (1997). Differences in perceived shape from shading correlate with activity in early visual areas. *Current*

Biology, *7*(2), 144–147. http://doi.org/10.1016/S0960-9822(06)00058-3

Imhof, E. (1967). Shading and Shadows. In *Cartographic Relief Representation* (Vol. 2007, pp. 159–212).

Jacobs, R. a. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences*, *6*(8), 345–350. http://doi.org/10.1016/S1364-6613(02)01948-4

Jenson, S. K. (1991). Applications of hydrologic information automatically extracted from digital elevation models. *Hydrological Processes*, *5*(October 1990), 31–44. http://doi.org/10.1002/hyp.3360050104

Kettunen, P., & Sarjakoski, T. (2009). Cartographic portrayal of terrain in oblique parallel projection. *Proc. 24th International Cartographic Conference*, (NOVEMBER 2009), 15–21. Retrieved from http://w.icaci.org/files/documents/ICC_proceedings/ICC2009/html/refer/20_10.pdf

Kleffner, D. a, & Ramachandran, V. S. (1992). On the perception of shape from shading. *Perception & Psychophysics*, *52*(1), 18–36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1635855

Kruijff, E., & Ii, J. E. S. (2010). Perceptual Issues in Augmented Reality Revisited, 3–12.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. J. (1995). Measurement and Modeling of Depth Cue, *35*(3), 389–412.

Langer, M. S., & Bülthoff, H. H. (2001). A prior for global convexity in local shape-from-shading. *Perception*, *30*(4), 403–10. http://doi.org/10.1068/p3178

Liu, B., Gould, S., & Koller, D. (2010). Single Image Depth Estimation From Predicted Semantic Labels, 1253–1260.

Liu, B., & Todd, J. T. (2004). Perceptual biases in the interpretation of 3D shape from shading. *Vision Research*, *44*(18), 2135–2145. http://doi.org/10.1016/j.visres.2004.03.024

Lobben, A. K. (2007). Navigational map reading: Predicting performance and identifying relative influence of map-related abilities. *Annals of the Association of American Geographers*, *97*(1), 64–85. http://doi.org/10.1111/j.1467-8306.2007.00524.x

Lovell, P. G., Bloj, M., & Harris, J. M. (2012). Optimal integration of shading and binocular disparity for depth perception. *Journal of Vision*, *12*(1), 1–1. http://doi.org/10.1167/12.1.1

Mamassian, P., & Goutcher, R. (2001). Prior Knowledge On The Illumination Position Brief article Prior knowledge on the illumination position, *81*(SEPTEMBER 2001), 1–9. http://doi.org/10.1016/S0010-0277(01)00116-0

Mamassian, P., Jentzsch, I., Bacon, B. A., & Schweinberger, S. R. (2003). Neural correlates of shape from shading. *NeuroReport*, *14*(7), 971–975. http://doi.org/10.1097/01.wnr.0000069061.85441.f2

Mamassian, P., & Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Research*, *38*(18), 2817–2832. http://doi.org/10.1016/S0042-6989(97)00438-0

Martin, D. W. (2008). *Doing Psychology Experiments* (Vol. 25). http://doi.org/10.1007/s13398-014-0173-7.2

Mehrabi, M., Peek, E. M., Wuensche, B. C., & Lutteroth, C. (2013). Making 3D Work: A Classification of Visual Depth Cues, 3D Display Technologies and Their Applications. *Proceedings of the Fourteenth Australasian User Interface Conference - Volume 139*, 91–100. Retrieved from http://dl.acm.org/citation.cfm?id=2525493.2525503

Meyer, M. (2015). Influence of Stereoscopic Displays on the Terrain Reversal Effect in Satellite Imagery.

Miller, H., Richardson, L., Sexton, N., Koontz, L., Loomis, J., & Koontz, S. (2012). The users, uses, and value of Landsat imagery in the United States. *2012 Socio-Economic Benefits Workshop: Defining, Measuring, and Communicating the Socio-Economic Benefits of Geospatial Information*. http://doi.org/10.1109/SeBW.2012.6292281

Morgenstern, Y., Murray, R. F., & Harris, L. R. (2011). The human visual system's assumption that light comes from above is weak. *Proceedings of the National Academy of Sciences*, *108*(30), 12551–12553. http://doi.org/10.1073/pnas.1100794108

Pentland, A. (1989). Shape information form shading: A theory about human perception.

Phillips, R. J., Lucia, A., & Skelton, N. (1975). Some Objective Tests of the Legibility of Relief Maps. *Cartographic Journal, The*, *12*(1), 39–46. http://doi.org/10.1179/000870475787338977

Pirscher, B., Foelsche, U., Lackner, B. C., & Kirchengast, G. (2007). Local time influence in single-satellite radio occultation climatologies from Sun-synchronous and non-Sun-synchronous satellites. *Journal of Geophysical Research Atmospheres*, *112*(11), 1–13. http://doi.org/10.1029/2006JD007934

Polys, N. F., Kim, S., & Bowman, D. A. (2005). Effects of Information Layout , Screen Size , and Field of View on User Performance in Information-Rich Virtual Environments, 46–55.

Ramachandran, V. S. (1988). Perceiving shape from shading.

Reichel, F. D., & Todd, J. T. (1990). Perceived depth inversion of smoothly curved surfaces due to image orientation. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 653–664. http://doi.org/10.1037/0096-1523.16.3.653

Řeřábek, M., Goldmann, L., Lee, J. S., & Ebrahimi, T. (2011). Motion parallax based restitution of 3D images on legacy consumer mobile devices. *MMSP 2011 - IEEE International Workshop on Multimedia Signal Processing*. http://doi.org/10.1109/MMSP.2011.6093789

Rogers, B., & Graham, M. (1979). Motion parallax as an independent cue for depth perception. *Perception*, *38*(6), 125–134. http://doi.org/10.1068/p080125

Rogers, B., & Graham, M. (1982). Similarities Between Motion in Human Depth Parallax and Perception *.

Sacks, O. (2006). Stereo Sue. *The New Yorker*, *82*(8), 64–73. Retrieved from http://www.superioreye.com/Reading for Residence Doctors/NewYorker Stereo Sue.pdf

Saraf, A. K., Das, J. D., Agarwal, B., & Sundaram, R. M. (1996). False topography perception phenomena and its correction. *International Journal of Remote Sensing*, *17*(18), 3725–3733.

http://doi.org/10.1080/01431169608949180

Saraf, A. K., Ghosh, P., Sarma, B., & Choudhury, S. (2005). Development of a new image correction technique to remove false topographic perception phenomena. *International Journal of Remote Sensing*, *26*(8), 1523–1529. http://doi.org/10.1080/0143116031000101648

Saraf, A. K., Zia, M., Das, J., Sharma, K., & Rawat, V. (2011). False topographic perception phenomena observed with the satellite images of Moon’s surface. *International Journal of Remote Sensing*, *32*(24), 9869–9877. http://doi.org/10.1080/01431161.2010.550950

Saraf, A., Sinha, S., Ghosh, P., & Choudhury, S. (2007). A new technique to remove false topographic perception phenomenon and its impacts in image interpretation. *International Journal of Remote Sensing*, *28*(5), 811–821. http://doi.org/10.1080/01431160701269796

Schiller, P. H., Slocum, W. M., Jao, B., & Weiner, V. S. (2011). The integration of disparity, shading and motion parallax cues for depth perception in humans and monkeys. *Brain Research*, *1377*, 67–77. http://doi.org/10.1016/j.brainres.2011.01.003

Sexton, I., & Surman, P. (1999). Stereoscopic and Autostereoscopic Display Systems - An In-Depth Review of Past, Present, and Future Technologies, (May), 85–99.

Sharma, K., Saraf, A. K., Das, J., & Baral, S. S. (2015). False topographic perception phenomena: observed in satellite images of Mars.

Smallman, H. S., Cook, M. B., Manes, D. I., & Cowen, M. B. (2007). NAÏVE REALISM IN TERRAIN APPRECIATION, 1–6.

Sun, J., & Perona, P. (1998). Where is the sun? *Nature Neuroscience*, 183–184. http://doi.org/10.1038/630

Tachikawa, T., Kaku, M., Iwasaki, A., Gesch, D., Oimoen, M., Zhang, Z., … Carabajal, C. (2011). by the ASTER GDEM Validation Team , with contributions by : Archive Center and the Joint Japan-US ASTER Science Team Table of Contents.

Taylor, J. C., Brewer, T. R., & Bird, a. C. (2010). Monitoring landscape change in the National Parks of England and Wales using aerial photo interpretation and GIS. *International Journal of Remote Sensing*, *21*(13–14), 2737–2752. http://doi.org/10.1080/01431160050110269

Todd, J. T., & Norman, J. F. (2003). The visual perception of 3-D shape from multiple cues: are observers capable of perceiving metric structure? *Perception & Psychophysics*, *65*(1), 31–47. http://doi.org/10.3758/BF03194781

Toppe, R. (1987). Terrain models-A tool for natural hazard mapping. *IAHS Publications*, (162). Retrieved from http://iahs.info/redbooks/a162/iahs_162_0629.pdf

Uratani, K., Machida, T., Kiyokawa, K., & Takemura, H. (2005). A Study of Depth Visualization Techniques for Virtual Annotations in Augmented Reality.

Van Beurden, M. H. P. H., Kuijsters, A., & IJsselsteijn, W. a. (2010). Performance of a path tracing task using stereoscopic and motion based depth cues. *2010 2nd International Workshop on Quality of Multimedia Experience, QoMEX 2010 - Proceedings*, 176–181. http://doi.org/10.1109/QOMEX.2010.5516268

Vezzani, S., Kramer, P., & Bressan, P. (2015). Stereokinetic Effect, Kinetic Depth Effect, and Structure from Motion. *Oxford Handbook of Perceptual Organization*, (November 2015). http://doi.org/10.1093/oxfordhb/9780199686858.013.053

Vuong, Q. C., Domini, F., & Caudek, C. (2006). Disparity and shading cues cooperate for surface interpolation. *Perception*, *35*(2), 145–155. http://doi.org/10.1068/p5315

Westheimer, G. (1994). The Ferrier Lecture, 1992. Seeing depth with two eyes: stereopsis., *257*(1349), 205–214.

Westheimer, G. (2011). Three-dimensional Displays and Stereo Vision. *Proceedings. Biological Sciences / The Royal Society*, *278*(1716), 2241–8. http://doi.org/10.1098/rspb.2010.2777

Willett, W., Jenny, B., Isenberg, T., Dragicevic, P., Willett, W., Jenny, B., … Dragicevic, P. (2015). Lightweight Relief Shearing for Enhanced Terrain Perception on Interactive Maps. *Proceedings of* *the 33rd ACM Conference on Human Factors in Computing Systems (CHI 2015)*. http://doi.org/10.1145/2702123.2702172

Wu, B., Li, H., & Gao, Y. (2013). Investigation and remediation of false topographic perception phenomena observed on Chang'E-1 lunar imagery. *Planetary and Space Science*, *75*, 158–166. http://doi.org/10.1016/j.pss.2012.10.018

Young, M. J., Landy, M. S., & Maloney, L. T. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Research*, *33*(18), 2685–2696. http://doi.org/10.1016/0042-6989(93)90228-O

Zhang, D., Zhu, W., Chen, X., Zheng, Z., Zhou, X., & Jiang, T. (2016). A correction technique for false topographic perception of remote-sensing images based on an inverse topographic correction technique. *International Journal of Digital Earth*, *8947*(September), 1–14. http://doi.org/10.1080/17538947.2016.1187672

Zhou, A., Zhang, X., & Gao, L. (2006). DEM terrain reversal. *Geography and Geo-Information Science*, *22*(6), 42–44.

Appendix

Overview Correction Methods

Criteria / Method	improved relief perception	keeps orientation (North reference)	works with rotation	keeps colour information	keeps classification	keeps texture	works for everybody	easy and fast	individually adaptable / interac-tion	Literature
180˚ rotation	++/--	--	-	++	++	++	+	++	--	Bernabé-Poveda et al. 2005; Gil et al. 2014; Saraf et al. 1996
negative of the image	+	+/-	--	--	--	+	+	++	--	Bernabà-Poveda et al. 2005; Bernabé-Poveda et al. 2011; Gil et al. 2014; Saraf et al 1996
enhanced color information	+/-	++	-	-	+/-	+	+/-	+	--	Bernabé-Poveda et al. 2011; Saraf et al. 2007;
SRM overlay	+	++	-	+ /-	+	+/-	+/-	+/-	+/-	Bernabé-Poveda et al. 2005; Bernabé-Poveda et al. 2011; Gil et al. 2014; Saraf et al. 2005; Wu et al. 2013
Stereopsis	+/-	+	+	-	-	+	--	-	-	Aubin & Arguin 2014; Hubona et al. 1999; Meyer 2015
Perspectives	+/-	+?	- ?	++	+	+	+ ?	--	--	Bernabé-Poveda et al. 2005; Toutin 1998
Elevation Infor-mation	--	++	--	++	++	++	++	++	+/-	Bernbaé-Poveda et al. 2005
Motion	+	+	- ?	+/-	+/-	+/-	?	-	+	Hubona et al. 1999; Willet et al. 2015

Stimuli Experiment I and II

Tasks (Experiment I and II):

TRE-questions:

The line between A, B and C appears as:

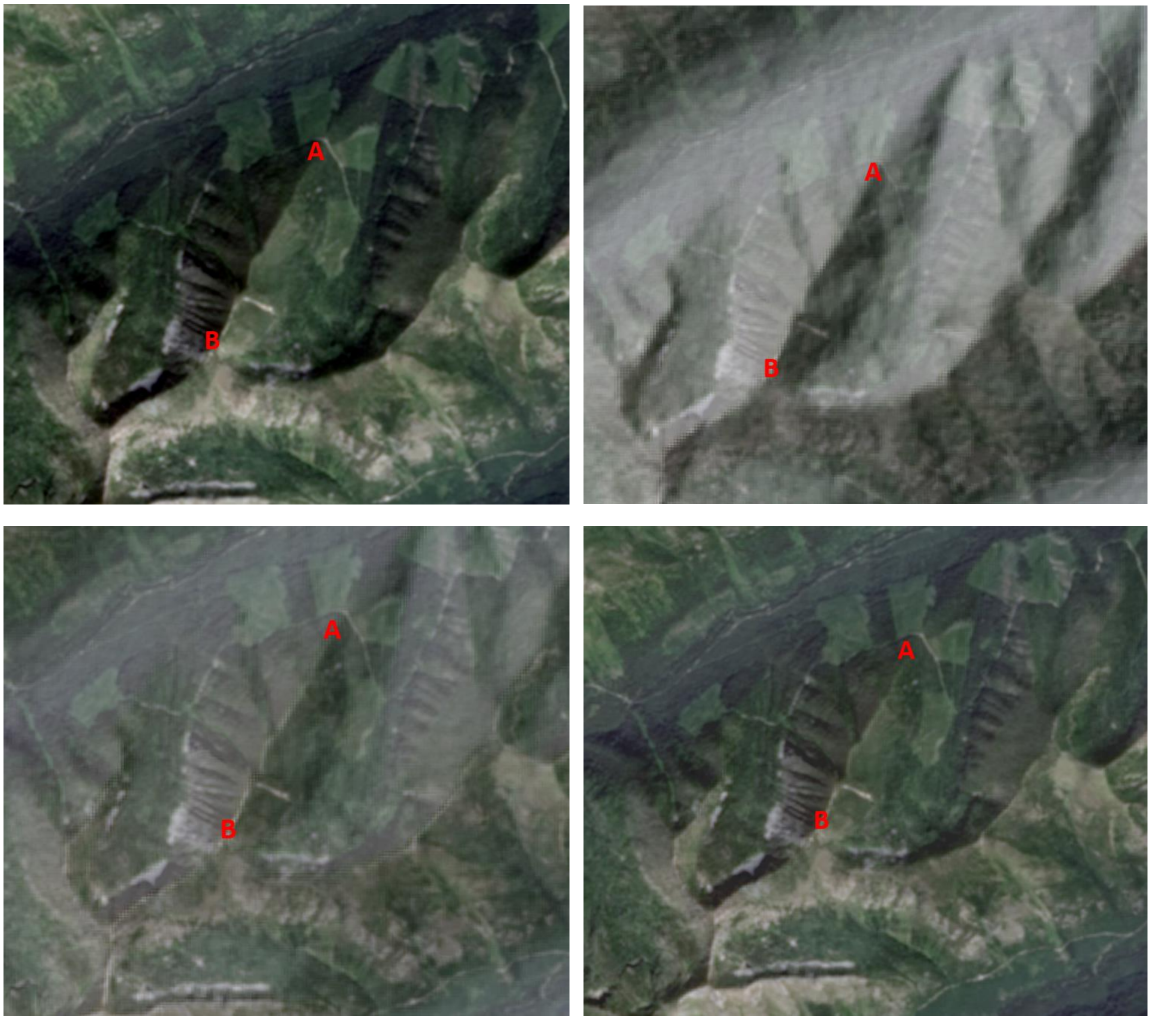
1 clearly a valley	2	3 ambiguous	4	5 clearly a ridge
○	○	○	○	○

LC-questions:

What do you see in the red square?
(It is possible to give more than one answer.)

- ☐ Forestland
- ☐ Grassland
- ☐ Rock / sand
- ☐ Snow / ice
- ☐ Water
- ☐ Ambiguous / not sure
- ☐ None of the above

Experiment I, TRE-questions (Examples)



Top left: Original

Top right: correction 45

Bottom left: correction 65

Bottom right: correction 85

Experiment I, LC-questions (Examples)

Top left: Original

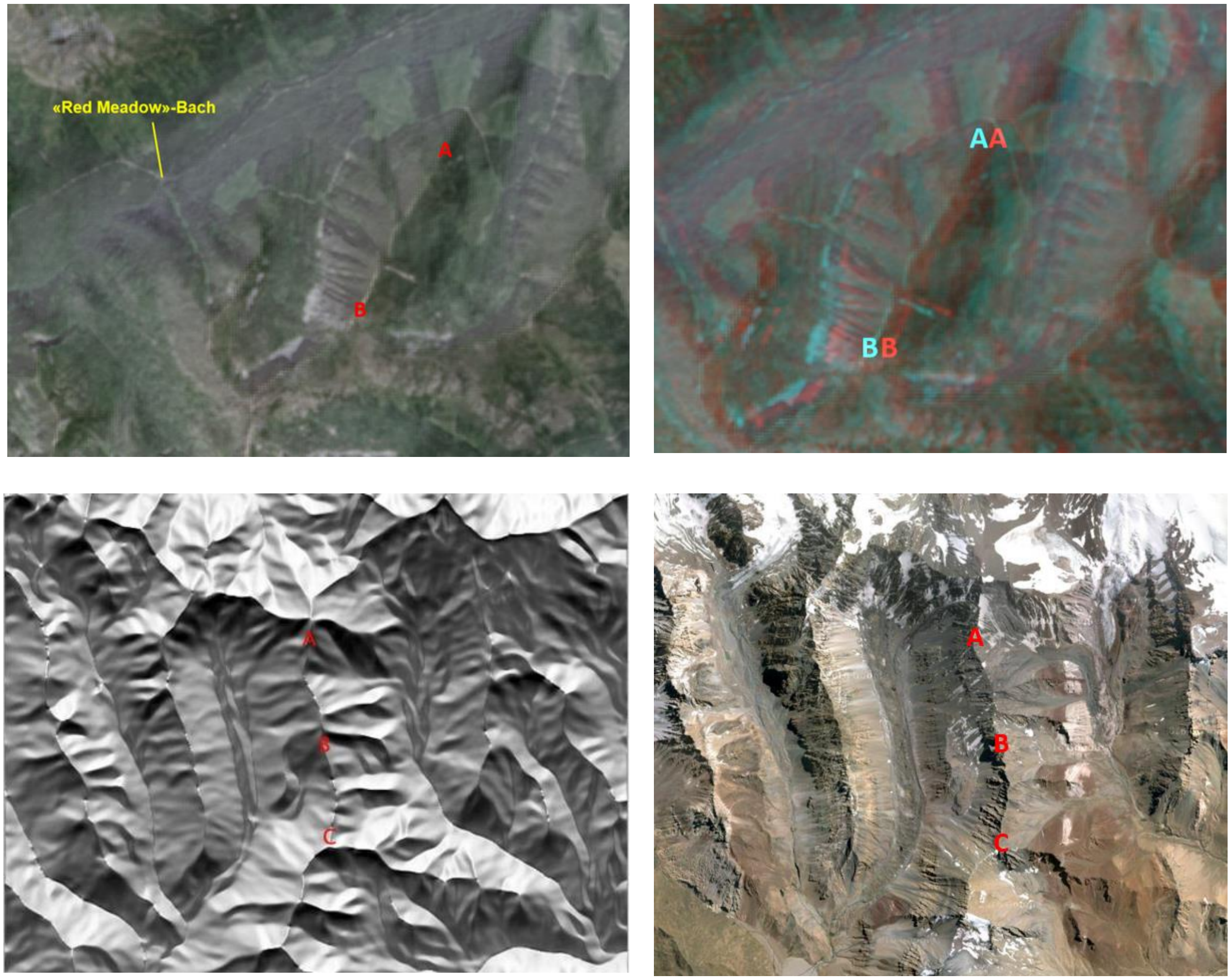
Top right: correction 45

Bottom left: correction 65

Bottom right: correction 85

Experiment II: TRE-questions (Examples)

The visualizations for ‘Original’ and ‘SRM-overlay’ (= correction 65 from Experiment I) are the same as in Experiment I.



Top left: Label

Top right: Stereo

Bottom left: SRM (Experiment II, Study 2)

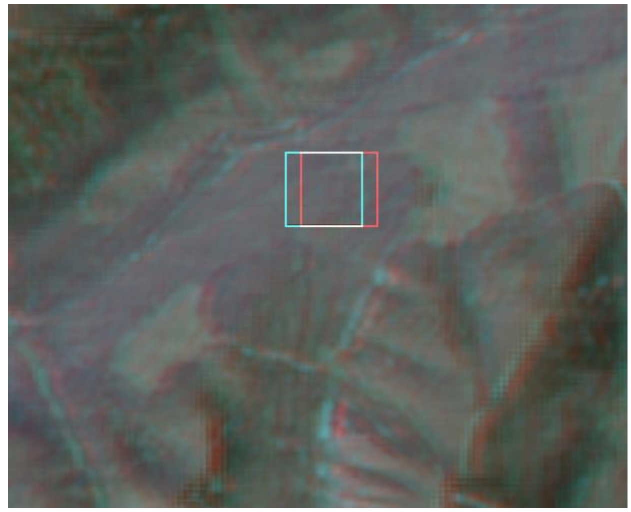
Bottom right: Satellite image corresponding to the SRM (Experiment II, Study 2)

An example of a motion image for TRE-questions is available on this website:

http://www.geo.uzh.ch/~ghartung/MA.html

Experiment II: LC-questions (Examples)

The visualizations for ‘Original’ and ‘SRM-overlay’ (= correction 65 from Experiment I) are the same as in Experiment I.



Left: Label

Right: Stereo

An example of a motion image for LC-questions is available on this website:

http://www.geo.uzh.ch/~ghartung/MA.html

Experiment I: Pre-Questionnaire

Geographic information visualization in satellite images and shaded reliefs

Welcome and thank you for accepting our invitation!

As part of a master thesis project in the Geographic Department at the University of Zurich we aim to enhance the visualization methods in the geo-information (GI) domain. This questionnaire is designed to gather data and opinions from a broad audience on selected topics from the field of geographic information visualization. The results are expected to help document how geographic visualizations are perceived and what influences our visual perception.

The participant information will be treated with full confidentiality and anonymity. Completing the questionnaire will take you about 20-25 minutes.

We are grateful for your participation.
If you would like to receive reports or papers resulting from this study, or for any other questions, please do not hesitate to contact us.

Gianna Hartung (gianna.hartung@uzh.ch) and Arzu Coltekin (arzu.coltekin@geo.uzh.ch)
http://www.geo.uzh.ch/~ghartung/
http://www.geo.uzh.ch/~arzu

Part I

The study consists of three parts. In this first part we are going to ask you different questions about yourself and your experience with certain topics.

Click "Next" to start the questionnaire.

About you

What is your gender?

- ◯ Male
- ◯ Female

What is your age?

[▾]

Education

What is the highest level of education you have completed?

- ◯ Basic education (less than high school diploma)
- ◯ High school diploma or equivalent
- ◯ University / College: Bachelor's degree or equivalent
- ◯ University / College: Master's degree or equivalent
- ◯ University / College: Doctoral degree

Language

Is English your first language?

- ◯ Yes
- ◯ No

English level

You marked that English is not your first language, please rate your English level:

1 Beginner	2	3 Intermediate	4	5 Proficiency
◯	◯	◯	◯	◯

Writing direction

Please indicate the direction in which you are writing most of the time.

- ◯ From left to right
- ◯ From right to left
- ◯ Both directions equally often
- ◯ Other (please specify)

[ ]

Hemisphere

On which hemisphere do you live most of the time?

- Northern hemisphere
- Southern hemisphere
- Both equally often
- Equatorial region

Map and image use

In your job: How often do you work with the following categories?

	Never	Rarely (approx. once a year)	Regularly (approx. once a month)	Often (approx. once a week)	Very often (approx. daily)
Cartography (maps)	○	○	○	○	○
Satellite imagery	○	○	○	○	○
3D(geo)visualizations (e.g. Google Earth, Street View)	○	○	○	○	○
Photography interpretation	○	○	○	○	○
Fine arts	○	○	○	○	○

In your leisure time: How often do you work with the following categories?

	Never	Rarely (approx. once a year)	Regularly (approx. once a month)	Often (approx. once a week)	Very often (approx. daily)
Cartography (maps)	○	○	○	○	○
Satellite imagery	○	○	○	○	○
3D(geo)visualizations (e.g. Google Earth, Street View)	○	○	○	○	○
Photography interpretation	○	○	○	○	○
Fine arts	○	○	○	○	○

Prescription glasses / contact lenses

Do you have prescription glasses and/or contact lenses?

- Yes
- No

Prescription glasses / contact lenses II

Do you use them now?

- Yes
- No

Color blindness

Have you ever been told by a professional that you have color blindness?

- Yes
- No

If you know, please note what type of color deficiency you have and how strong it is. If you do not know, click "Next" to continue.

Handedness

I am

1 totally left-handed	2	3 ambiguous	4	5 totally right-handed
○	○	○	○	○

Mask

Here you see a video with two equal rotating masks. Look at the nose of one of the masks (it does not matter which one). Does the nose always appear convex (pops out) when the rotating mask looks at you directly?
Click "Next" as soon as you answered the question.

- Yes, the nose always appears convex (pops out) during the rotation when the mask looks directly at me. The nose might appear distorted when the mask looks to the side (left or right).
- Sometimes the nose appears convex (pops out), sometimes it seems to be concave (goes in) depending on which side of the mask I see during the rotation.

Sleep

How many hours did you sleep last night?

- Less than 3 hours
- 3 hours
- 4 hours
- 5 hours
- 6 hours
- 7 hours
- 8 hours
- More than 8 hours

Experiment I: Post-Questionnaire

Part III

You're almost there! We will now move on to the last section. In this part we are going to ask you different questions about the tasks you solved before.

Click "Next" to start the questionnaire.

Tasks

Did you find the tasks boring or tiring?

1 Not at all boring / tiring	2	3 Sometimes boring / tiring	4	5 Very boring / tiring
○	○	○	○	○

Preference Sat

To identify land forms (hills, valleys): How do you rate the quality of the images above?

Very good	Good	Acceptable	Bad	Very bad
○	○	○	○	○

To identify land cover (e.g. vegetation, snow, rock): How do you rate the quality of the images above?

Very good	Good	Acceptable	Bad	Very bad
○	○	○	○	○

Preference S65

To identify land forms (hills, valleys): How do you rate the quality of the images above?

Very good	Good	Acceptable	Bad	Very bad
○	○	○	○	○

To identify land cover (e.g. vegetation, snow, rock): How do you rate the quality of the images above?

Very good	Good	Acceptable	Bad	Very bad
○	○	○	○	○

Preference S85

To identify land forms (hills, valleys): How do you rate the quality of the images above?

Very good	Good	Acceptable	Bad	Very bad
○	○	○	○	○

To identify land cover (e.g. vegetation, snow, rock): How do you rate the quality of the images above?

Very good	Good	Acceptable	Bad	Very bad
○	○	○	○	○

Preference S45

To identify land forms (hills, valleys): How do you rate the quality of the images above?

Very good	Good	Acceptable	Bad	Very bad
○	○	○	○	○

To identify land cover (e.g. vegetation, snow, rock): How do you rate the quality of the images above?

Very good	Good	Acceptable	Bad	Very bad
○	○	○	○	○

Contradiction

Have you ever noticed a contradiction between landform (valley, hill) and landcover (e.g. snow, vegetation, river)?

- ○ Yes
- ○ No

Contradiction II

Please explain briefly the observed contradictions.

How did you answer the question then?

- ☐ Based on my perception
- ☐ Based on my interpretation
- ☐ I cannot remember

Change of landform

In any of the images, did you perceive a landform first as a valley, but shortly after it appeared as a ridge (or vice versa)?

- ○ Yes
- ○ No

Change of landform II

Overall you saw 80 images. Estimate how often such a change of the landform occured.

- ○ In less than 10 images (less than 10% of the images)
- ○ In more than 10 images (more than 10% of the images)
- ○ In more than 20 images (more than 25% of the images)
- ○ In more than 40 images (more than 50% of the images)
- ○ In more than 60 images (more than 80% of the images)

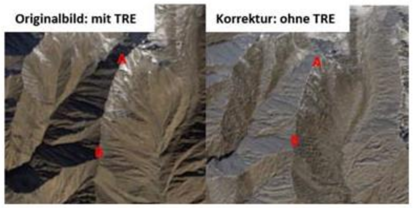
How did you answer the questions in such a case?

- ☐ Based on my first impression
- ☐ Based on my last impression
- ☐ I guessed the answer
- ☐ I don't know

Terrain reversal effect

In this study we have been testing the terrain reversal effect (TRE), a perceptual phenomenon where landforms appear inverted, e.g. valleys are perceived as ridges and vice versa (see image below). Did you notice this phenomenon?

Yes	No
○	○



Participation

Did you participate in a similar study before?

- ○ Yes
- ○ No
- ○ I cannot remember

The end

You now arrived at the end of the study.

Thank you very much for taking the time to participate in this study. If you have any comments, please note them here. If you'd like to hear from us regarding the study results, or if you wish that we get back to you for a question you may have, please leave us your email address. Note that this is optional and we will treat your information with utmost confidentiality.

Experiment II: Consent Form

Universität Zürich - **Teilnehmerinformation und Einwilligungsformular** **Evaluation von geographische Visualisierungen mit Satellitenbilder: Eine Fallstudie mit Augenbewegungsanalyse**
Mai / Juni 2016
Teilnehmernummer:

Zweck der Studie

Sie sind eingeladen, an einer Studie zur Evaluation von geographischen Visualisierungen mit Satellitenbildern und Einfluss der Händigkeit darauf teilzunehmen. Wir möchten dabei Informationen über die Gestaltung und Benutzerfreundlichkeit von digitalen Karten gewinnen.

Ablauf der Studie und damit verbundene Risiken

Nach Ihrem definitiven Entscheid, an der Studie teilzunehmen, füllen Sie zuerst einen kurzen Fragebogen aus, in dem Sie unter anderem Angaben zu Ihrer Person machen. Im Anschluss daran werden Sie gebeten, einige Aufgaben am Computer zu lösen. Dazu benützen Sie vorgegebene digitale Bilder. Währenddessen wird Ihre Interaktion mit dem Computer mit Hilfe einer Kamera, eines Mikrofons und eines Blickregistrierungssystems aufgezeichnet. Das Blickregistrierungssystem ermöglicht es, Ihre Augenbewegungen ohne jeglichen Körperkontakt aufzuzeichnen. Dazu wird nicht sichtbares Licht im nahen Infrarotbereich verwendet, das keine unangenehmen Auswirkungen hat. Nach der Aufzeichnung werden Sie einen weiteren Fragebogen ausfüllen.

Die Studie dauert ungefähr 60 Minuten und beinhaltet keinerlei Risiken für Sie.

Vertraulichkeit der Daten

Jegliche Information, die während der Studie in Verbindung mit Ihnen gebracht werden kann, wird vertraulich behandelt und nur mit Ihrer ausdrücklichen Erlaubnis an Dritte weitergegeben. Mit Ihrer Unterschrift erlauben Sie uns, die Ergebnisse des Versuchs mehrmals zu publizieren. Dabei werden keinerlei persönliche Informationen veröffentlicht, die es ermöglichen, Sie zu identifizieren.

Abfindung

Wir bieten keine finanzielle Entschädigung für die Teilnahme an der Studie an. Auch Kosten, die Ihnen für die Teilnahme an der Studie entstehen, werden nicht rückerstattet.

Bekanntgabe der Ergebnisse

Wenn Sie über die Ergebnisse der Studie auf dem Laufenden gehalten werden möchten, bitten wir Sie, dem Versuchsleiter oder der Versuchsleiterin Ihre E-Mail-Adresse und/oder Anschrift zu hinterlassen. Eine Kopie der Publikation(en) wird Ihnen daraufhin zugestellt.

Einwilligung

Ihre Entscheidung, an der Studie teilzunehmen oder nicht, wird etwaige zukünftige Beziehungen mit der Universität Zürich nicht beeinträchtigen. Entscheiden Sie sich dafür, an der Studie teilzunehmen, steht es Ihnen jederzeit frei, die Teilnahme ohne Begründung abzubrechen.

Sollten Sie Fragen haben, zögern Sie bitte nicht, uns diese zu stellen. Sollten zu einem späteren Zeitpunkt Fragen aufkommen, wenden Sie sich bitte an Gianna Hartung (079 537 32 77, giannahartung@gmx.ch) oder Dr. Arzu Coltekin (044 635 54 40, arzu@geo.uzh.ch).

Sie erhalten eine Kopie dieses Dokuments.

Universität Zürich - **Teilnehmerinformation und Einwilligungsformular** Evaluation von geographische Visualisierungen mit Satellitenbilder: Eine Fallstudie mit Augenbewegungsanalyse
Mai / Juni 2016
Teilnehmernummer:

Mit Ihrer Unterschrift bestätigen Sie, obenstehende Informationen gelesen und verstanden zu haben und willigen ein, unter den dort beschriebenen Bedingungen am Experiment teilzunehmen.

…………………………………… ……………………………………

Unterschrift des Teilnehmers | Unterschrift des Experimentleiters

…………………………………… ……………………………………

Vor- und Nachname in Blockschrift | Vor- und Nachname in Blockschrift

……………………………………

Ort / Datum

Universität Zürich - **Teilnehmerinformation und Einwilligungsformular** Evaluation von geographische Visualisierungen mit Satellitenbilder: Eine Fallstudie mit Augenbewegungsanalyse
Mai / Juni 2016
Teilnehmernummer:

WIDERRUF DER EINWILLIGUNG

Hiermit möchte ich meine Einwilligung, an der oben beschriebenen Studie teilzunehmen, widerrufen.

…………………………………… ……………………………………

Unterschrift | Ort / Datum

……………………………………

Vor- und Nachname in Blockschrift

Mit dem Widerruf der Einwilligung beeinträchtigen Sie in keiner Weise Ihre Beziehungen mit der Universität Zürich. Der Widerruf kann jederzeit und ohne Angabe von Gründen beantragt werden.

Den Widerruf der Einwilligung bitte an *Dr. Arzu Coltekin, Geographische Informationsvisualisierung und Analyse, Departement für Geographie, Universität Zürich, Winterthurerstrasse 140, 8057 Zürich* senden.

Experiment II: Pre-Questionnaire

Teil I_Pre

Teilnehmernummer (wird von der Studienleitung ausgefüllt)

Geschlecht, Alter

Geschlecht:

- Männlich
- Weiblich

Alter:

Ausbildung

Welches ist Ihre höchste abgeschlossene Ausbildung?

- Keine
- Real- oder Sekundarstufe
- Lehrabschluss
- Kantonsschule / Gymnasium / Fachmaturitätsschule
- Fachhochschule / Höhere Fachschule / Pädagogische Hochschule
- Universität: Bachelor
- Universität: Master
- Universität: Doktor

Karten und Bilder

Wie oft nutzen Sie eine oder mehrere der folgenden Kategorien?

	Nie	Selten (jährlich)	Regelmässig (monatlich)	Oft (wöchentlich)	Sehr oft (täglich)
Kartographie (Karten)	○	○	○	○	○
Satellitenbilder	○	○	○	○	○
3D-Geovisualisierungen (z. Bsp. Google Earth, Street View)	○	○	○	○	○
Interpretation von Fotographien oder Bildern (z. Bsp. in Museen)	○	○	○	○	○

Computer

Wie oft nutzen Sie einen Computer oder Laptop?

Nie	Selten	Regelmässig	Täglich
○	○	○	○

Brille / Kontaktlinsen

Benutzen Sie eine Brille oder Kontaktlinsen?

- Ja
- Nein

Brille / Kontaktlinsen II

Tragen Sie diese jetzt?

- Ja, Brille
- Ja, Kontaktlinsen
- Nein

Farbenblindheit

Wurde bei Ihnen eine Farbenblindheit diagnostiziert?

- ◯ Ja
- ◯ Nein

Bitte geben Sie an, welche Art von Farbenblindheit Sie haben und wie stark diese ist. Falls Sie dies nicht wissen, klicken Sie auf "Weiter".

[ ]

Händigkeit

Ich bin

1 Linkshänder	2	3 beidhändig	4	5 Rechtshänder
◯	◯	◯	◯	◯

Maske

Hier sehen Sie eine Videosequenz mit zwei rotierenden Masken. Schauen Sie auf die Nase einer Maske (es spielt keine Rolle, welche Maske Sie anschauen).
Sieht die Nase immer konvex aus (springt aus dem Bild), wenn die rotierende Maske Sie direkt anschaut?
Bitte beantworten Sie die Frage gemäss Ihrer Wahrnehmung.

Klicken Sie auf "Weiter" sobald Sie die Frage beantwortet haben.

- ◯ Ja, die Nase sieht immer konvex aus (springt aus dem Bild) währen der Rotation und wenn die Maske mich direkt anschaut. Die Nase sieht vielleicht etwas verzerrt aus, wenn die Maske zur Seite schaut (links oder rechts).
- ◯ Manchmal sieht die Nase konvex aus (springt aus dem Bild), manchmal sieht die Nase konkav aus (geht ins Bild hinein). Das hängt davon ab, welche Seite der Maske ich gerade sehe während der Rotation.

Schlaf

Wie viele Stunden haben Sie letzte Nacht geschlafen?

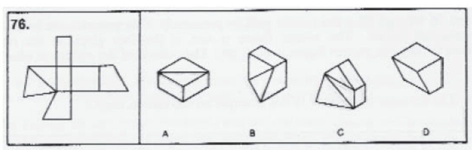
- ◯ Weniger als 3 Stunden
- ◯ 3 Stunden
- ◯ 4 Stunden
- ◯ 5 Stunden
- ◯ 6 Stunden
- ◯ 7 Stunden
- ◯ 8 Stunden
- ◯ Mehr als 8 Stunden

Experiment II: Spatial Ability Test

SDT_Teilnehmernummer

Teilnehmernummer (wird von der Studienleitung ausgefüllt)

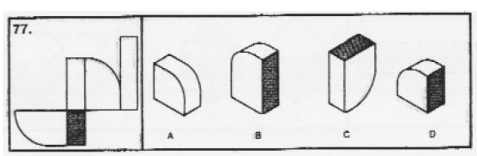
1



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- A
- B
- C
- D

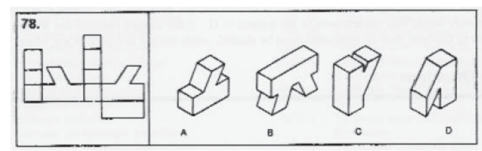
2



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- A
- B
- C
- D

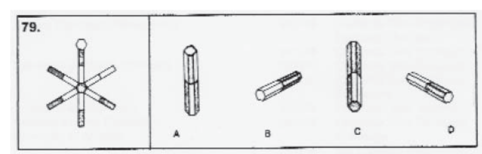
3



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- A
- B
- C
- D

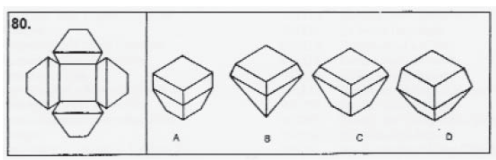
4



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- A
- B
- C
- D

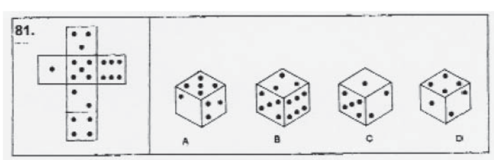
5



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- A
- B
- C
- D

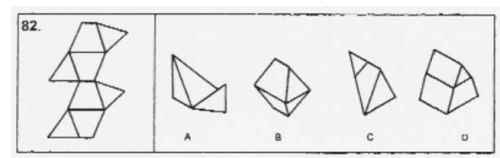
6



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- A
- B
- C
- D

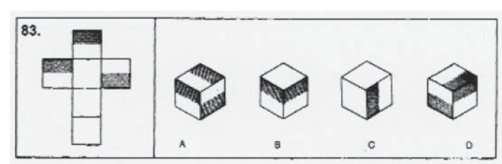
7



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- A
- B
- C
- D

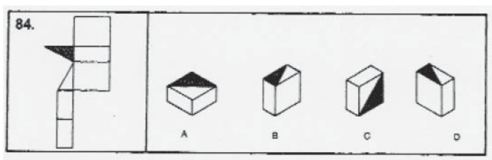
8



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- ◯ A
- ◯ B
- ◯ C
- ◯ D

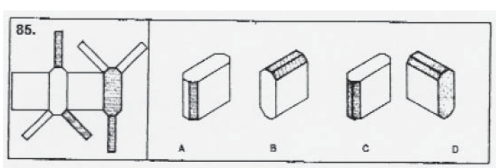
9



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- ◯ A
- ◯ B
- ◯ C
- ◯ D

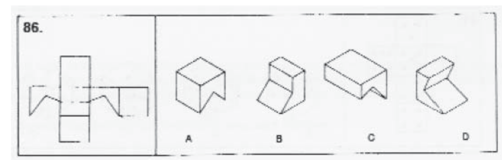
10



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- ◯ A
- ◯ B
- ◯ C
- ◯ D

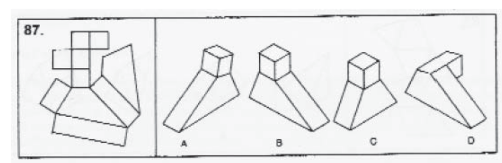
11



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- ◯ A
- ◯ B
- ◯ C
- ◯ D

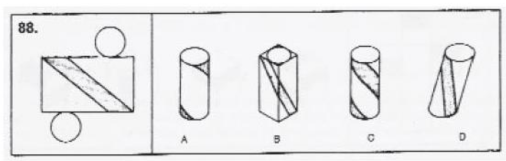
12



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- A
- B
- C
- D

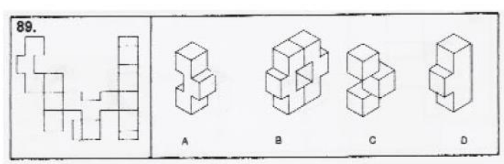
13



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- A
- B
- C
- D

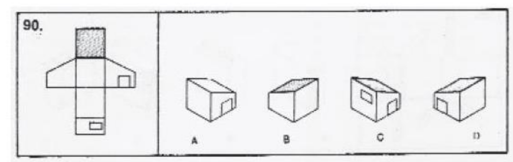
14



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- A
- B
- C
- D

15



Welches Objekt passt zur aufgefalteten Version auf der linken Seite?

- A
- B
- C
- D

Experiment II: Stereoscopic Vision Test (Handout)

Stereopsistest

Im Verlauf des Experiments wird Ihr räumliches Sehen getestet. Dafür brauchen wir folgende Angaben von Ihnen:

Teilnehmernummer: ___________

Grösse (in cm): ___________

Alter: ___________

Brille oder Kontaktlinsen (Ja/Nein): ___________

Setzen Sie sich vor den Laptop mit einem Abstand von einer Armlänge. Setzen Sie die 3D-Brille auf und starten Sie den Test. Im Test sehen Sie Stereogramme, in welchen ein Objekt entweder vor oder hinter dem Rest des Bildes erscheint. Sehen Sie sich das Bild ein paar Sekunden an und klicken Sie dann auf die für Sie richtig erscheinende Antwort.

Der Test ist in Englisch verfasst. Sie haben folgende Antwortmöglichkeiten:

- The square is "**in front of**" the screen. = Die Box ist **vor** dem Bildschirm.
- The square is "**behind**" the screen. = Die Box ist **hinter** dem Bildschirm.
- «**not sure**» = **nicht sicher**

Bitte halten Sie den vor dem Test festgelegten Abstand zum Monitor (eine Armlänge) während des ganzen Tests ein!

stereo acuity: ___________

Experiment II: Post-Questionnaire

Teil 3_Post

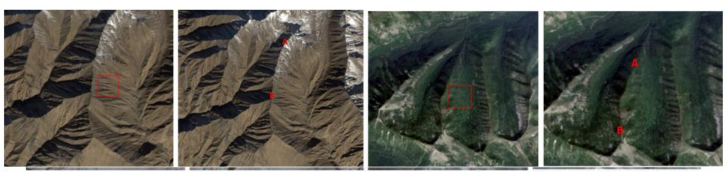
Teilnehmernummer (wird von der Studienleitung ausgefüllt)

Aufgaben

Waren die Aufgaben langweilig oder ermüdend?

1 Überhaupt nicht langweilig / ermüdend	2	3 Manchmal langweilig / ermüdend	4	5 Sehr langweilig / ermüdend
○	○	○	○	○

Präferenz original



Um Landformen zu erkennen (Hügel, Täler): Wie bewerten Sie die Qualität der obigen Bilder?

Sehr gut	Gut	Akzeptabel	Schlecht	Sehr schlecht
○	○	○	○	○

Um die Landbedeckung zu erkennen (z. Bsp. Vegetation, Schnee, Gestein etc.): Wie bewerten Sie die Qualität der obigen Bilder?

Sehr gut	Gut	Akzeptabel	Schlecht	Sehr schlecht
○	○	○	○	○

Präferenz corronline

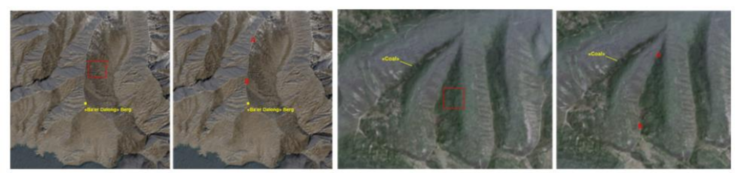
Um Landformen zu erkennen (Hügel, Täler): Wie bewerten Sie die Qualität der obigen Bilder?

Sehr gut	Gut	Akzeptabel	Schlecht	Sehr schlecht
○	○	○	○	○

Um die Landbedeckung zu erkennen (z. Bsp. Vegetation, Schnee, Gestein etc.): Wie bewerten Sie die Qualität der obigen Bilder?

Sehr gut	Gut	Akzeptabel	Schlecht	Sehr schlecht
○	○	○	○	○

Präferenz label



Um Landformen zu erkennen (Hügel, Täler): Wie bewerten Sie die Qualität der obigen Bilder?

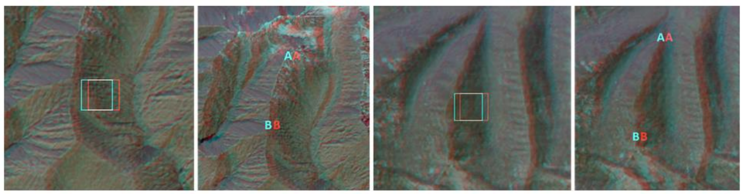
Sehr gut	Gut	Akzeptabel	Schlecht	Sehr schlecht
○	○	○	○	○

Um die Landbedeckung zu erkennen (z. Bsp. Vegetation, Schnee, Gestein etc.): Wie bewerten Sie die Qualität der obigen Bilder?

Sehr gut	Gut	Akzeptabel	Schlecht	Sehr schlecht
○	○	○	○	○

Präferenz stereo

Bitte setzen Sie die Anaglyphenbrille nochmals auf, nachdem Sie die Fragen gelesen haben.



Um Landformen zu erkennen (Hügel, Täler): Wie bewerten Sie die Qualität der obigen Bilder?

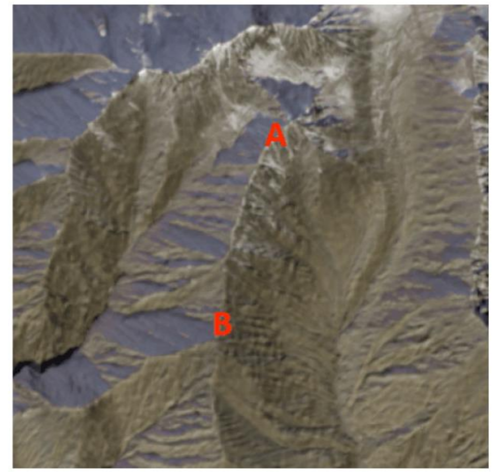
Sehr gut	Gut	Akzeptabel	Schlecht	Sehr schlecht
○	○	○	○	○

Um die Landbedeckung zu erkennen (z. Bsp. Vegetation, Schnee, Gestein etc.): Wie bewerten Sie die Qualität der obigen Bilder?

Sehr gut	Gut	Akzeptabel	Schlecht	Sehr schlecht
○	○	○	○	○

Sie können die Anaglyphenbrille nach Beantwortung der Fragen wieder absetzen.

Präferenz motion



Um Landformen zu erkennen (Hügel, Täler): Wie bewerten Sie die Qualität des obigen Bild?

Sehr gut	Gut	Akzeptabel	Schlecht	Sehr schlecht
○	○	○	○	○

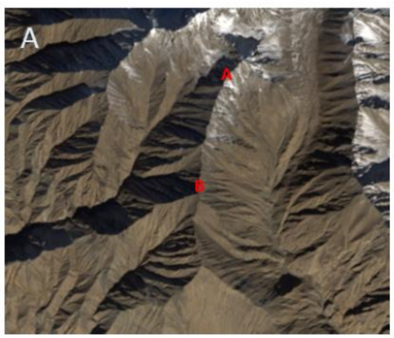
Um die Landbedeckung zu erkennen (z. Bsp. Vegetation, Schnee, Gestein etc.): Wie bewerten Sie die Qualität des obigen Bild?

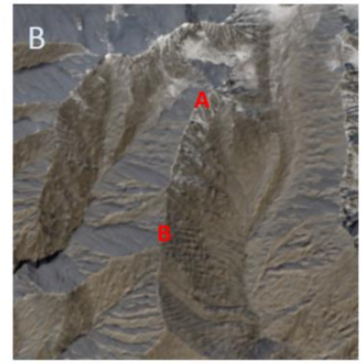
Sehr gut	Gut	Akzeptabel	Schlecht	Sehr schlecht
○	○	○	○	○

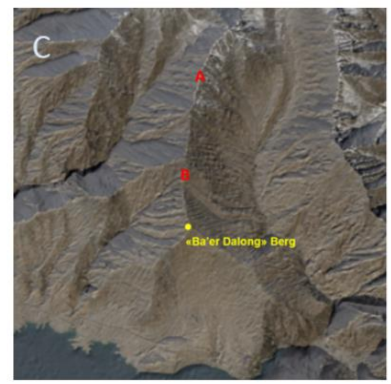
Gesamtpräferenz

Sie sehen unten jeweils ein Beispiel einer Visualisierungsart (Bild A - E), welche in der Studie verwendet wurde. Welche Visualisierung bevorzugten Sie bei der Beantwortung der Fragen? Bitte weisen Sie den Buchstaben A - E jeweils eine Nummer zu, so dass 1 = "beste Visualisierung" und 5 = "schlechteste Visualisierung".

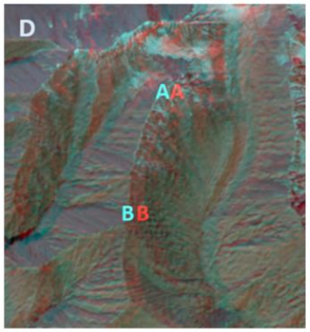
- A
- B
- C
- D
- E

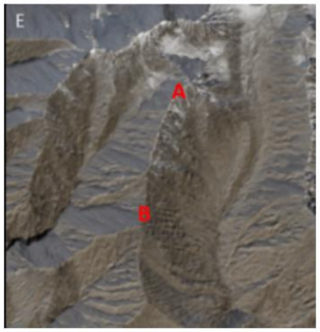






Setzen Sie nur für das folgende Bild "D" die Anaglyphenbrille nochmals auf:





Anaglyphenbrille

Haben Sie ein Unwohlsein verspürt, als Sie die stereoskopischen Bilder mit der Anaglyphenbrille betrachtet haben?

1 kein / wenig Unwohlsein	2	3 mittleres Unwohlsein	4	5 starkes Unwohlsein
○	○	○	○	○

Motion

Haben Sie ein Unwohlsein verspürt, als Sie die sich bewegenden Bilder angeschaut haben?

1 kein / wenig Unwohlsein	2	3 mittleres Unwohlsein	4	5 starkes Unwohlsein
○	○	○	○	○

Widerspruch

Haben Sie während der Studie jemals einen Widerspruch zwischen Landform (Tal, Berg) und Landbedeckung (z. Bsp. Schnee, Vegetation, Gewässer) bemerkt?

- ○ Ja
- ○ Nein

Widerspruch II

Bitte beschreiben Sie kurz diese Widersprüche.

Wie haben Sie die Fragen, in denen Sie Widersprüche bemerkt haben, beantwortet?

- ☐ Gemäss meiner Wahrnehmung
- ☐ Gemäss meiner Interpretation
- ☐ Ich weiss es nicht

Umkehrung der Landform

Haben Sie in einem der Bilder zuerst ein Tal wahrgenommen, kurz darauf aber sah die Landform aus wie ein Hügel (oder umgekehrt)?

- ○ Ja
- ○ Nein

Änderung der Landform II

Insegesamt haben Sie 148 Bilder gesehen. Schätzen Sie, wie oft eine Umkehrung der Landform vorkam.

- In weniger als 10 Bildern
- In ca. 10 - 20 Bildern
- In ca. 20 - 50 Bildern
- In ca. 50 - 100 Bildern
- In mehr als 100 Bildern

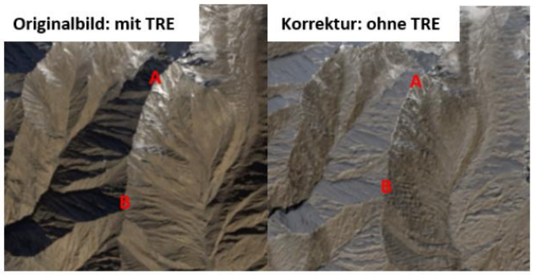
Wie habe Sie in diesem Fall die Frage beantwortet?

- Gemäss meines ersten Eindrucks
- Gemäss meines letzten Eindrucks
- Ich habe geraten
- Ich weiss nicht

Terrain reversal effect

In dieser Studie wurde der sogenannte "terrain reversal effect (TRE)" getestet. Der TRE ist ein visuelles Phänomen, bei dem Landformen umgekehrt erscheinen, z. Bsp. werden Täler als Hügel wahrgenommen und umgekehrt (s. Beispiel unten). Haben Sie dieses Phänomen bemerkt während der Beantwortung der Aufgaben?

- Ja
- Nein



Teilnahme

Haben Sie an einer ähnlichen Studie bereits zuvor teilgenommen?

- Ja
- Nein
- Ich weiss nicht

Ende

Sie sind jetzt am Ende der Studie angekommen. Vielen herzlichen Dank für Ihre Teilnahme.

Notieren Sie allfällige Bemerkungen bitte hier:

Falls Sie über die Studienresultate benachrichtigt werden möchten, geben Sie bitte Ihre E-Mail-Adresse an. Diese Angabe ist optional. Ihre E-Mail-Adresse wird nicht an Dritte weiteregegeben oder für andere Zwecke gebraucht.

Experiment II: Study Protocol

Material: Anaglyphenbrille, Einverständniserklärung, Handout Stereotest, Stift, Laptop, Aufladekabel für Laptop, Computermaus, Schokolade

Vorbereitung:

- Laptop: Pre-, Postquestionnaire, Spatial ability test und Stereotest (http://3d.mcgill.ca/cbc/) und Beispielbilder für Stereotest öffnen; Computer: TobiiStudio starten
- Teilnehmernummer überall einfüllen; Bildschirmauflösung einstellen (1600x1024)
- Licht auf maximale Helligkeit; Schild vor Türe umkehren

1. Begrüssung

- Stereotest Handout ausfüllen
- Einverständniserklärung abgeben oder lesen und ausfüllen
- Ablauf der Studie erklären: 3 Fragebögen am Laptop – Hauptteil am Computer – Schluss am Laptop, ca. 60 Minuten
- Fragen?

2. Pre-Questionnaire

Währenddessen: Daten von Stereo-Test Handout abschreiben

3. Surface Development Test

«Sie haben 6 Minuten Zeit, um so viele Fragen wie möglich zu beantworten. Sollten Sie die Antwort auf eine Frage nicht wissen, können Sie ausnahmsweise auch auf «Weiter» klicken.»

4. Stereotest

Anleitung dem Teilnehmenden zur Lektüre geben, währenddessen Voreinstellung für Test vornehmen.

«Der Test ist in Englisch. Bei sprachlichen Problemen, stehe ich Ihnen zur Verfügung. Setzen Sie sich vor den Laptop und strecken Sie einen Arm aus, so dass Sie mit den Fingerspitzen den Laptop-Bildschirm berühren können. Diesen Abstand sollten Sie während des ganzen Tests einhalten.»

- Anaglyphenbrille (gelb / blau) dem Teilnehmenden geben und Beispielbilder zeigen

Nach dem Test: Stereo acuity Wert abschreiben.

5. Hauptteil

Eye Tracker anstellen

«Wir starten nun mit dem Hauptteil der Studie am Computer. Hier werden Ihre Augenbewegungen mit einem Eyetracker aufgenommen. Dabei wird gemessen, wo Sie wie lange auf den Bildschirm schauen. Währen der Studie läuft eine Kamera mit Ton. Bitte achten Sie währen der Studie darauf, dass Sie sich nicht allzu viel bewegen und auch die Distanz zum Computer einhalten. Währen des Tests brauchen Sie die Anaglyphenbrille nochmals. Sie werden während des Tests darauf aufmerksam gemacht. Bitte bewegen Sie sich möglichst wenig, wenn Sie die Brille anziehen. Wenn Sie sich während des Tests unwohl fühlen sollten, melden Sie sich bitte. Beantworten Sie die Fragen so schnell und so genau wie möglich und danach, was sie auf dem Bild sehen.»

Kalibration durchführen & Start recoding

6. Post-Questionnaire

Anaglyphenbrille zur Seite legen!

7. Ende

Schoggi & Kopie consent form übergeben

Daten speichern und Backup durchführen

Personal Declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

Gianna Hartung

Zurich, September 30th 2016