Slovak University of Technology in Bratislava

Faculty of Informatics and Information Technologies

FIIT-10894-5737

Ing. Patrik Polatsek

MODELLING OF HUMAN VISUAL ATTENTION

Dissertation thesis

Degree Course:	Applied Informatics
Field of study:	9.2.9 Applied Informatics
Place of development:	Institute of Computer Engineering and Applied Informatics, FIIT STU Bratislava
Supervisor:	Assoc. Prof. Vanda Benešová

May 2019

Candidate: Ing. Patrik Polatsek
Institute of Computer Engineering and Applied Informatics
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava

Supervisor: Assoc. Prof. Vanda Benešová
Institute of Computer Engineering and Applied Informatics
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava

Reviewers: Assoc. Prof. Elena Šikudová
Department of Software and Computer Science Education
Faculty of Mathematics and Physics
Charles University

Prof. Pavel Slavík
Department of Computer Graphics and Interaction
Faculty of Electrical Engineering
Czech Technical University in Prague

Keywords: visual attention, saliency model, egocentric attention, motion saliency, color saliency, depth saliency, shape saliency, emotions, visualizations

Categories and Subject Descriptors (ACM Classification 1998)

I.2.10 [ARTIFICIAL INTELLIGENCE]: Vision and Scene Understanding – *Perceptual reasoning*
I.4.8 [IMAGE PROCESSING AND COMPUTER VISION]: Scene Analysis – *Color, Depth cues, Motion, Shape*

ANOTÁCIA

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný odbor: APLIKOVANÁ INFORMATIKA

Autor:	Ing. Patrik Polatsek
Dizertačná práca:	Modelovanie ľudskej vizuálnej pozornosti
Vedúci dizertačnej práce:	doc. Ing. Vanda Benešová, PhD.

máj, 2019

V posledných desaťročiach sa modelovanie vizuálnej pozornosti stalo dôležitou výskumnou oblasťou. Aby sme mohli simulovať ľudskú pozornosť, je potrebné do takéhoto modelu zakomponovať rôzne mechanizmy pozornosti, ktoré sú riadené stimulmi a cieľmi pozorovateľa. Kvôli zložitosti tohto procesu je potrebné skúmať nápadnosť jednotlivých faktorov, ktoré ovplyvňujú našu pozornosť, separátne.

V tejto práci sme skúmali ako príznaky na nízkej a strednej úrovni ako farba, pohyb, hĺbka a tvar ovplyvňujú vizuálnu pozornosť v našich vlastných experimentoch so sledovaním pohľadu. Na zmeranie týchto vplyvov sme použili viacero existujúcich ako aj vlastných modelov predpovedajúcich nápadnosť konkrétneho príznaku. Kvôli nedostatku datasetov so sledovaním pohľadu, ktoré by sa špecializovali na zvolenú vizuálnu nápadnosť, sme fixačné dáta z našich experimentov verejne sprístupnili.

Aby sme lepšie porozumeli procesu selektívnej pozornosti pri každodenných úlohách, vykonali sme tiež experimenty v reálnom prostredí, ktoré sme zaznamenávali z pohľadu prvej osoby. Naše výsledky ukázali silnú individuálnosť egocentrickej pozornosti, ktoré sa odlišuje od prezerania 2D obrazov, čiastočne kvôli binokulárnym podnetom, ktoré obohacujú vnímanie pozorovateľa. Preto odporúčame, aby boli použité špecializované modely nápadnosti pre egocentrické videnie. Napokon sme zistili, že vysoko-úrovňové faktory ako emócie alebo úlohy vykonávané na vizualizáciách taktiež ovplyvňujú pohľad človeka.

ANNOTATION

Slovak University of Technology in Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: APPLIED INFORMATICS

Author:	Ing. Patrik Polatsek
Dissertation thesis:	Modelling of Human Visual Attention
Supervisor:	Assoc. Prof. Vanda Benešová

2019, May

In recent decades, visual attention modelling became a prominent research area. In order to simulate human attention, a computational model has to incorporate various stimulus-driven and goal-directed attention mechanisms. Because of the complexity of this process, it is important to investigate saliency of each factor that affects our attention individually.

In this thesis we explored how low- and mid-level features such as color, motion, depth and shape influence visual attention in our own eye-tracking experiments. To measure these effects, we utilized various state-of-the-art as well as novel computational models which estimate saliency of a specific feature. Because of a lack of eye-tracking databases that would specialize primarily on selected feature saliency, we made fixation data from our experiments publicly available.

In order to deeper understand the process of selective attention in everyday actions, we conducted several experiments in real environments recorded from the first-person perspective. Our results showed that egocentric attention is very individual and differs from 2D image viewing conditions, partially due to binocular cues that enhance viewer's perception. We therefore suggest to employ specialized saliency models for egocentric vision. Finally, we found out that high-level factors such as individual's emotions and task-based analysis of visualizations influence human gaze behavior too.

Declaration of Honour

Hereby I declare that I wrote this thesis independently under professional supervision of Assoc. Prof. Vanda Benešová with citated bibliography.

May, 2019 in Bratislava signature

Acknowledgement

I would like to thank my supervisor Assoc. Prof. Vanda Benešová for her willingness, continual support and helpful advice during the work on this thesis. I also thank my family and friends for their unceasing encouragement and support.

Contents

1 Introduction **1**

1.1 Motivation 2

1.2 Challenges 3

1.3 Contribution of the Thesis 4

1.4 Structure of the Thesis 5

2 Related Work **7**

2.1 Human Visual System 8

2.2 Eye Movements 10

2.3 Feature Binding 10

2.3.1 Computational Models 12

2.4 Color Stimulus 15

2.4.1 Computational Models 16

2.5 Motion Stimulus 16

2.5.1 Computational Models 17

2.6 Depth Stimulus 18

2.6.1 Computational Models 20

2.7 Shape Stimulus 20

2.7.1 Computational Models 21

2.8 Emotional Factors 22

2.8.1 Computational Models 23

2.9 Visual Attention in Visualizations 23

2.9.1 Computational Models 24

2.10 Evaluation Metrics of Computational Models 24

2.10.1 Location-Based Metrics 24

2.10.2 Distribution-Based Metrics . . . 25
2.10.3 Human Inter-Observer . . . 26

3 Contribution **27**
3.1 Egocentric Motion Saliency Modelling . . . 27
3.1.1 Motivation . . . 27
3.1.2 Analyzed Computational Model . . . 28
3.1.3 Evaluation . . . 29
3.1.4 Summary . . . 31
3.2 Visual Attention to Color . . . 31
3.2.1 Motivation . . . 32
3.2.2 Experimental Study on Color Saliency . . . 32
3.2.3 Experimental Results and Discussion . . . 33
3.2.4 Summary . . . 35
3.3 Visual Attention to Shape . . . 35
3.3.1 Motivation . . . 35
3.3.2 Experimental Study on Shape Saliency . . . 36
3.3.3 Proposed Computational Models . . . 38
3.3.4 Experimental Results and Discussion . . . 40
3.3.5 Summary . . . 44
3.4 Visual Attention to Egocentric Depth . . . 45
3.4.1 Motivation . . . 45
3.4.2 Experimental Study on Depth Saliency . . . 45
3.4.3 Experimental Results . . . 48
3.4.4 Discussion . . . 50
3.4.5 Summary . . . 53
3.5 Static Feature-Based Egocentric Visual Attention . . . 53
3.5.1 Egocentric Experiment . . . 54
3.5.2 Experimental Results and Discussion . . . 57
3.5.3 Summary . . . 63
3.6 Emotionally-Tuned Visual Attention . . . 64
3.6.1 Motivation . . . 65
3.6.2 Emotion-Based Visual Attention Experiment . . . 65

3.6.3 Experimental Results 68
3.6.4 Discussion 71
3.6.5 Summary 72
3.7 Visual Attention during Task-Based Analysis of Information Visualizations 72
3.7.1 Motivation 72
3.7.2 Memorability Experiment by Borkin et al. 73
3.7.3 Task-Based Visual Analysis Experiment 73
3.7.4 Experimental Results 79
3.7.5 Discussion 84
3.7.6 Summary 86

4 Conclusions **87**

A DVD Contents **105**

B Resumé **107**

C Publications of the Author with Internal Categorization and Relevant Citations **117**

Chapter 1

Introduction

Visual attention is a set of cognitive processes that selects relevant information and filters out irrelevant information from the environment [13]. Therefore, it plays an important role in the control of head and eye movements. Scene scan is performed by a sequence of rapid movements called *saccades* and *fixations*, as shown in Figure 1.1. During a fixation, the eye is relatively still to acquire visual information from the focus of interest [181, 75].

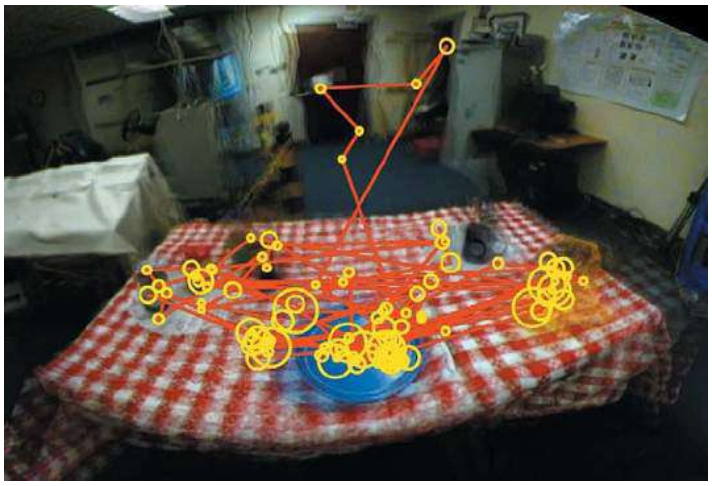


Figure 1.1: Trace of fixations denoted by yellow circles whose size is proportional to the fixation duration and saccades denoted by red lines [181].

Attention is influenced by both *bottom-up* (stimulus-driven) and *top-down* (goal-directed, knowledge-based) factors. Bottom-up attention is very rapid and directly affected by environmental stimuli. It was originally developed in the brain to monitor the environment for critical, potentially dangerous stimuli. It is associated with the so-called "*pop-out effect*" when low-level salient visual features perceptually stand out from their neighborhood, such as intensity, color, texture and motion contrasts. Much slower top-down attention enhances stimulus-driven attention. It is influenced by cognitive factors such as individual's knowledge, expectations, goals and tasks. Psychologists assume that bottom-up and top-down processing works together to organize and interpret visual information from the environment. This is referred to as *perception* [25, 74, 115, 57].

Visual saliency has been researched in many research areas including psychology, neurobiology, image processing and computer vision [16]. There are two different approaches how

to define saliency of an image [148] (see example in Figure 1.2):

1. We can *measure saliency* based on human behavior when viewing an image, e.g. by recording fixation data using eye-trackers.
2. We can also *predict saliency* by computational saliency models. They compute a *saliency map* from an image, representing a topographical map of conspicuousness [16].

(a) Image. (b) Fixation heatmap. (c) Saliency map [96].

Figure 1.2: Saliency defined by human fixations and by a computational saliency model [14].

Visual saliency modelling can be applied in many areas of computer science including image processing [2, 150, 7], computer graphics [47], advertising [156], robotics [28, 183], surveillance systems [52, 129] and human-computer interaction [182, 30].

1.1 Motivation

Most standard bottom-up saliency models define a saliency map as a global conspicuity measure of a location which combines information of multiple individual maps representing a local conspicuity of a particular feature dimension [116]. Therefore, it is necessary to separately explore local feature conspicuity first to reduce the remaining gap to human gaze behavior.

To approach human-level performance, saliency models should implement mechanisms that lead to object perception such as perceptual feature grouping, contour and shape detection [13]. However, attention affected by object shape and size is seldom integrated in computational models.

Existing benchmarks show that computational saliency has made a significant progress in human gaze prediction [18, 14, 225][1], especially thanks to neural network architectures [89, 123]. Since selection of visual information from the environment is a complex process, state-of-the-art saliency models can cover only a limited number of factors that affect attention. Even though the gap between predicted saliency maps and human fixations is reduced for some image databases, the results also indicate that their accuracy relies on a category of a scene [14, 130]. For instance, a model presented by Judd et al. [105] is tuned to predict attention in natural outdoor scenes employing high-level features detection such as human faces, horizons and cars that are the key salient objects for this scene category. However, adding high-level features can lead to ignore low-level saliency. In addition, objects may not be equally relevant to observers, but the model does not take the relative importance of detected objects into consideration [32, 130]. Top-performing neural network models are pre-trained

[1] `http://saliency.mit.edu`

on a large image database for object recognition. Subsequently, the models are retrained on human fixations to predict saliency. Such models can therefore excel in object recognition for a major scene category in a training dataset (including objects like humans, animals, roads, cars) [178]. For example, a deep neural network proposed by Huang et al. [89] which computes saliency with high precision for natural images, fails to determine regions that attract attention in case of artificial images. Hence, to improve saliency computation, new specialized image databases are required for training [32].

Top-down attention is influenced not only by prior knowledge. Most models have been evaluated against fixations from free viewing. To simulate human vision system, a computational model should be also enhanced by given tasks that guide attention [189].

Because of the increasing usage of wearable cameras, attention modelling should be focused on videos from the first-person perspective [151]. Saliency could help to interpret daily activities in egocentric videos in terms of health, social interaction analysis, traffic security, or in market research. In contrast to on-screen view, egocentric vision is also affected by object distance from an observer, object and observer's motion [75].

1.2 Challenges

The primary goal of this thesis is to individually study various aspects of visual attention using novel eye-tracking experiments and computational saliency models. Since the majority of these factors has been explored marginally so far, we recorded fixation data in image viewing conditions and real environments to deeper understand human visual system and increase the performance of saliency models.

The main challenges of the thesis are summarized as follows:

1. **Egocentric visual attention:** Attention in real environments recorded by miniaturized wearable cameras, such as GoPro and Google Glass, differs from attention when viewing 2D video clips. In addition to feature saliency estimated by conventional models, saliency modelling for videos from the first-person perspective should also focus on factors that influence egocentric vision including dynamic ego-motion, binocular depth information and observer's prior knowledge about the environment [75].

2. **Visual attention to 2D static visual features:** Bottom-up models usually employ intensity, color and orientation contrasts to identify salient regions of a scene [25, 75, 16]. Color is one the most important visual feature in the environment, used in perceptual feature grouping and object discrimination [74, 57]. Although color processing is utilized in most saliency models, such as opponent process theory implemented in Itti et al.'s model [96], more research in color perception is needed to understand how color contrasts and psychological color associations, for instance red color associated with danger, affect human fixations [51]. Furthermore, there is also object-based attention affected by object size and shape [44]. Shape saliency and perception is a complex process that should be modelled on a local, contour level as well as global, object shape level [142, 13].

3. **Task-driven visual attention in visualizations:** A majority of saliency models have been evaluated during task-free viewing of natural scenes. Saliency modelling has been already used as a quality metric for visualizations [10]. However, there are some

notable differences between natural images and classic charts used in information visualization. In addition, there is little evidence how much influence bottom-up saliency has on task-based visual analysis. According to the Guided Search Model [208], both bottom-up and top-down information are weighted according to the task.

4. **Impact of actual individual's mood on visual attention and visual search:** Broaden-and-build theory states that positive emotions enhance visual attention [62]. While some saliency models incorporate detection of emotional content in affective scenes, they have been never evaluated using emotionally-neutral stimuli under a particular emotional state of observers.

1.3 Contribution of the Thesis

This dissertation contributes to the area of visual attention and saliency modelling. This work explores various bottom-up and top-down aspects of attention. The principal contributions of the thesis are:

1. **Exploring egocentric saliency in a real environment** (published in [234], [237] and [233]):
   - *Analysis of egocentric motion saliency and observer's surprise using an egocentric dataset*: In the author's master thesis, an egocentric saliency model which integrates spatial and motion contrasts and motion memory was introduced [234, 237]. In contrast, the dissertation thesis evaluates the prediction ability of this model on a larger egocentric dataset, compares with other computational models and discusses the importance of saliency from motion contrasts and surprise. This work therefore partially covers Challenge 1.
   - *Analysis of egocentric depth saliency using own eye-tracking experiment and novel computational models, creating a fixation dataset for egocentric depth saliency*: We performed an experiment with eye-tracking glasses to explore the relationship between stereoscopic depth and saliency and made it publicly available. We analyzed the effects of relative depth and depth contrasts on human visual attention. We used recorded fixations to propose novel methods to to estimate depth saliency based on depth contrast or depth weighting. The whole study is related to Challenge 1.
   - *Analysis of static feature-based attention using own eye-tracking experiment and novel computational models, creating a fixation dataset for egocentric static saliency*: We conducted a novel eye-tracking experiment to find out how all low- and mid-level static features influences egocentric vision including the depth effects, in contrast to prior research. We therefore evaluated stat-of-the-art and novel saliency models to measure the effects of feature saliency such as intensity, color, orientation, depth, shape and center-bias. Because of lack of the first-person perspective datasets, we recorded own video sequences with depth information and fixation data using eye-tracking glasses and Kinect device. We made this dataset publicly available for model evaluation. This experiment is related to Challenge 1 and partially to Challenge 2.
2. **Exploring the effect of input factors such as color, shape and emotional situation**

of user in image viewing conditions and their incorporation into computation saliency models (published in [235]):

- *Analysis of shape saliency using own eye-tracking experiment and novel computation models, creating a fixation dataset for shape saliency*: To explore shape saliency separately with aim to minimize the effect of other attention aspects, we conducted an eye-tracking experiment with own images that contain only object silhouettes. Data from our experiment are open to the public. We proposed 30 own techniques to compute shape saliency based on commonly used shape descriptors and matchers. Our models detect either local contour saliency or global shape rarity. This study is related to Challenge 2.
- *Analysis of color saliency using own eye-tracking experiment, creating a fixation dataset for color saliency*: We conducted an eye-tracking experiment using own images designed to maximize the influence of color on attention to investigate color saliency individually. We made data from this experiment publicly available. To cover Challenge 2, we focused on the effect of relative color saliency related to color differences and the effect of absolute color importance related to learned color associations.
- *Analysis of emotionally-tuned visual attention using own eye-tracking experiment and evaluation of a computational model, creating a fixation dataset affected by emotions*: We performed an eye-tracking experiment focused on human emotions. We induced positive or neutral mood to participants and record their fixations during various tasks. We made data from this experiment publicly available. We analyzed the effect of induced emotions on visual attention and visual search and discussed their differences to cover Challenge 4.

3. **Exploring attention during task-based visual analysis in visualizations** (published in [236]):
   - *Analysis of goal-directed visual attention in visualizations using own eye-tracking experiment and evaluation of computational models, creating a fixation dataset for task-based visual analysis*: We performed a task-based eye-tracking experiment with information visualizations and compared the fixation data with existing memorability experiment with same images to investigate to study the effect of bottom-up and top-down attention when performing a task. Recorded fixation data are publicly available. Eventually, we evaluated state-of-the-arts models against these datasets to find out if they may be also applied in visualizations [10]. This experiment covers Challenge 3.

1.4 Structure of the Thesis

The remainder of this work is organized into three chapters. Chapter 2 introduces elementary principles of visual attention and perception and discussed state-of-the-art saliency models. Chapter 3 presents the contribution of this thesis – evaluation of an egocentric motion saliency model (Section 3.1), a color saliency experiment (Section 3.2), a shape saliency experiment with computational shape models (Section 3.3), a binocular depth saliency experiment with computational depth models (Section 3.4), an egocentric experiment focused on various static salient features (Section 3.5), an emotionally-tuned experiment (Section 3.6)

and a task-based visual analysis experiment (Section 3.7). And finally, Chapter 4 summarizes proposed computational models and findings of the eye-tracking experiments.

Chapter 2

Related Work

Even though the visual attention research is most concerned with the selective attention, there are various kinds of attention with different abilities which are defined as follows [130]:

1. *focused attention*: the ability to focus on a target stimulus,
2. *sustained attention*: the ability to focus on an activity over an extended period of time,
3. *selective attention*: the ability to select specific stimuli while ignoring others,
4. *alternating attention*: the ability to shift a focus of attention,
5. *divided attention*: the ability to focus on multiple tasks at the same time.

Recent decades of visual attention research have brought many computational models that can be grouped by various criteria, including [16, 130]:

- **factors that attract attention:** *bottom-up* (stimuli-driven) factors, *top-down* (goal-directed, knowledge-based) factors and their combination,
- **temporality:** *spatial* models based solely on a current scene and *temporal* models based on the accumulated prior knowledge, motion analysis or combination of both aspects,
- **stimuli type:** *static* stimuli such as intensity, color and depth and *dynamic* stimuli such as motion and flicker,
- **task type:** *free viewing*, *visual search tasks* and other more *complex tasks* (e.g. driving),
- **saliency units:** *location-based* models that assign saliency values to each location defined by pixels and macro-blocks and *object-based* models that either extract salient objects from location-based saliency or directly compute saliency at object level.
- **size of information** used in saliency estimation: *local* information based only on a subregion of an image, *global* information based on a whole image.

There are many methods to predict human visual attention. This chapter groups the methods according to aspect of visual attention they use in saliency computation and introduces basic principles of human visual system.

2.1 Human Visual System

Vision, the most important human sense, is based on the visible spectrum of light reflected from objects into eyes. Visible light is a small range of electromagnetic spectrum from about 400 to 700 nm [75].

Human visual system starts when light enters the eye covered by the *cornea* through a small hole in the center of the *iris* called the *pupil*. It is projected onto a thin light-sensitive layer called the *retina* containing millions of photoreceptors that turn light energy into electrical signals (see Figure 2.1) [108].

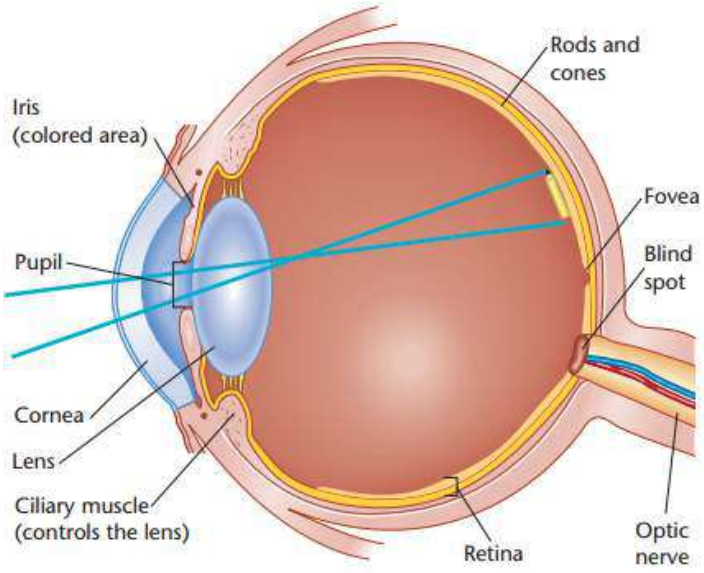


Figure 2.1: Structure of the eye [108].

There are two types of photoreceptors – *rods* (about 120 millions) and *cones* (about 6 millions). Due to the sensitivity to dim light, rods are responsible for night vision, but they are unable to distinguish color. They are not useful under daylight, because bright light bleach them. They are sensitive to changes in brightness, but not to changes in wavelength. Their peak spectral sensitivity is about 500 nm. In contrast to rods, the major function of cones is bright-light and color vision. Cones respond faster to light and their sensitivity to light intensity is lower than rods. There are *S-cones*, *M-cones* and *L-cones* with peak sensitivity of 420 nm, 530 nm and 560 nm wavelengths, respectively (Figure 2.2) [13].

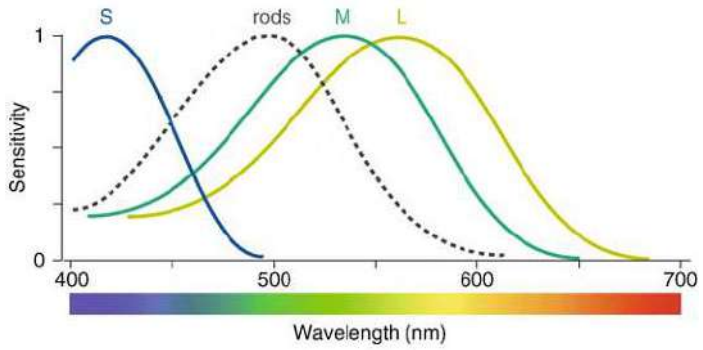


Figure 2.2: Sensitivity of photoreceptors [13].

The distribution of photoreceptors varies in the retina. A small central area of the retina called the *fovea* has the greatest density of photoreceptors. It is specialized for the detailed vision and contains only cones (1% of all cones in the retina). After a direct look at an object, its image falls on the fovea. The rest of the retina is called the *peripheral retina* and consists of both receptor types. Towards the periphery, there are more and more rods converging onto bipolar cells, due to which peripheral vision has low sensitivity to spatial location and high sensitivity to low spatial frequency (dim light) and high temporal frequency stimuli [75, 13, 108].

A typical ganglion cell has a receptive field with *center-surround* organization, as shown in Figure 2.3. A group of ganglions which are excited when light hits their center and inhibited when it hits the surroundings are called *on-center* cells. On the other hand, *off-center* cells are excited and inhibited in reverse order [108, 75].

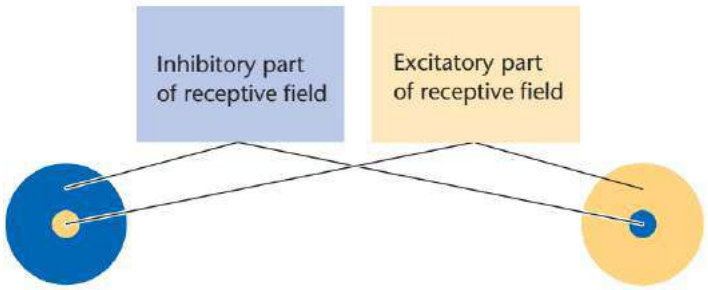


Figure 2.3: Two types of receptive fields of ganglion cells – on-center (left) and off-center (right) [108].

Ganglion cells fall into the following three primary categories [108]:

1. *Parvocellular* neurons can be found primarily in the fovea. They have small cell bodies and receptive fields. Their function is detailed and color analysis of stationary objects.

2. *Magnocellular* neurons located throughout the retina have larger cell bodies and receptive fields. They respond strongly to movement and broad shapes of objects, but without color sensitivity.

3. *Koniocellular* neurons are also located throughout the retina with various functions for analysis of visual information.

Figure 2.4 illustrates a route of visual signals. Axons of ganglion cells form the left and right *optic nerves* that meet at the optic chiasm. Most axons continue to the *lateral geniculate nucleus* (LGN) in the thalamus whose cells have similar organization to ganglion cells. Another axons proceed to the *superior colliculus* which is responsible for eye movements and other visual behavior. Visual signals from the LGN finally go to area V1 in the *primary visual cortex* and then spread out to other cortical areas [232, 75, 57].

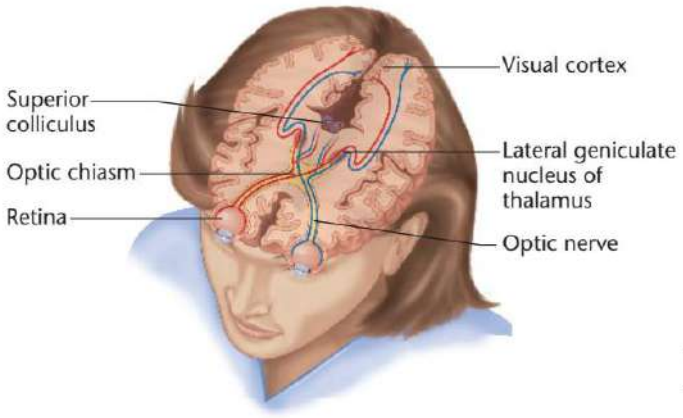


Figure 2.4: Route of visual signals [108].

2.2 Eye Movements

To perceive the environment, human eyes perform various voluntary and involuntary movements that fall into three main categories [74][1]:

1. **Compensatory eye movements** are automatic reflexes that compensate head motions to keep incoming image on the retina relatively still.
2. **Saccadic eye movements** quickly move the fovea from one point of interest to another, usually three times per second. Most saccades are not controlled by vision and take approximately from 20 up to 40 msec. Exploration of a visual scene occurs during short pauses between saccades are referred to as *fixations* which last 50 to 600 msec. Even though the eye is relatively stationary during a fixation, it performs multiple micro-movements (tremor, micro-saccades and drifts).
3. **Vergence eye movements** control the registration of retinal images from both eyes.

In addition, there are other eye movements such as *maintained fixation* which occurs when eyes fixate a stationary object for a period of time and *smooth pursuit* referring to tracking of a moving object.

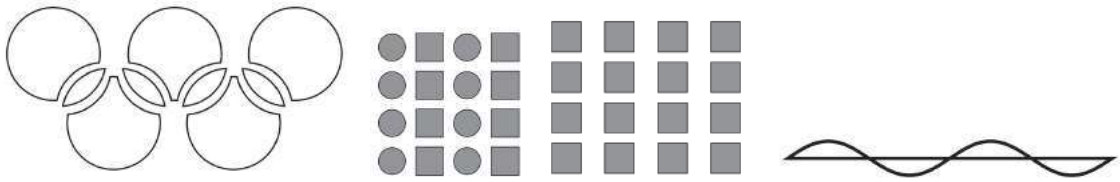
2.3 Feature Binding

The function of attention is also to solve the spatial *binding problem*. It refers to the combination of features to coherent visual perception [195]. Current behavioral studies have suggested two types of visual attention – *space-based* and *object-based* attention, which coexist in the visual system and may affect one another [159].

An early attempt to explain feature grouping and object perception was proposed by **Gestalt** psychologists. They proposed a number of perceptual principles called Gestalt laws (see examples in Figure 2.5) [207, 117, 118].

[1] https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/types-of-eye-movements/

The *law of Prägnanz*, i.e. the law of good figure, is the fundamental Gestalt law. According to the law, the simplest possible organization of the visual field is perceived. The *law of similarity* states that objects with similar visual features, such as color, orientation, texture, size and shape are perceived as a group. According to the *law of continuity*, we prefer perceptions of continuous objects to disjoint object parts. The *law of proximity* states that objects near each other are grouped together. According to the *law of common fate*, objects that move in the same direction are seen as being together. The *figure-ground segregation* explains how objects are separated from the background. According to this principle, we organize the visual field into stimuli that stand out (figure) and less important background (ground) [74, 57, 155, 232, 75].



(a) Prägnanz. (b) Similarity. (c) Proximity. (d) Continuity.

Figure 2.5: Examples of Gestalt laws. The first image is perceived as five circles. Objects in the second image are grouped into columns of squares and circles. The third image is seen as four columns of squares. The last image is perceived as two continuous lines [75, 232, 74].

Treisman and colleagues [196, 197, 195] defined the most famous theory about attention and feature binding called the **Feature Integration Theory** (FIT). The theory suggests two stages of object processing. During the first, *preattentive stage* visual features such as color and orientation are detected in the brain automatically without attention. The so-called preattentive features are encoded in parallel by a set of *feature maps*. Each map registers the activity in response to a particular feature but does not give information about spatial location. Attention is used in the second, *focused attention stage* to combine individual features. It scans serially a *master map of locations* by a spatial window and finds a correct combination of feature maps to form an object (see Figure 2.6). Therefore, the second stage is much slower in comparison to the parallel preattentive stage [203, 75, 25].

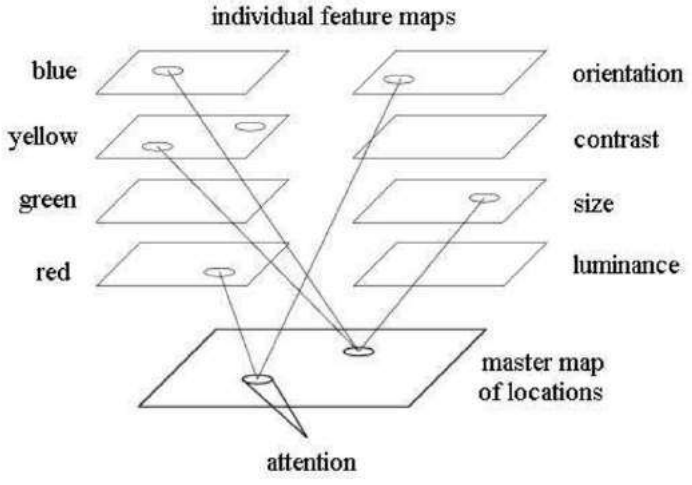


Figure 2.6: Feature Integration Theory [197, 203].

Koch and Ullman [116] extended the theory and presented the first architecture of selective attention. They defined a topographic *saliency map* that directly represents the overall conspicuity at each location. The map combines local conspicuity of a particular visual feature encoded in a feature map. Selective visual attention is always shifted to the next most salient location implemented using the *winner-take-all* (WTA) neural network which detects the location with highest saliency and *inhibition-of-return* (IOR) which suppresses saliency of the last attended location [95].

Beside FIT, there is another important theory of attention – the *Guided Search Model* proposed by Wolfe [210, 208, 211, 209]. It states that bottom-up attention is modulated by top-down attention depending on a performed visual search task (see Figure 2.7).

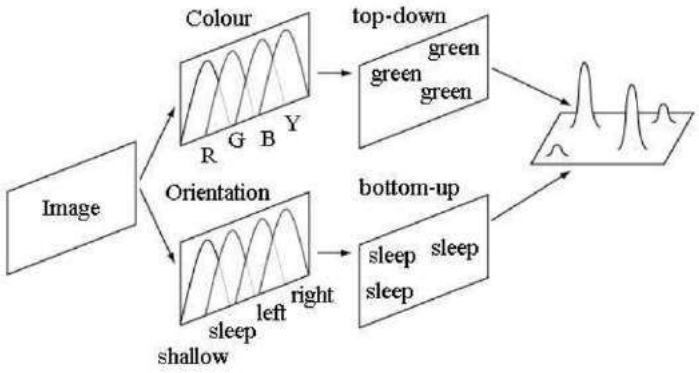
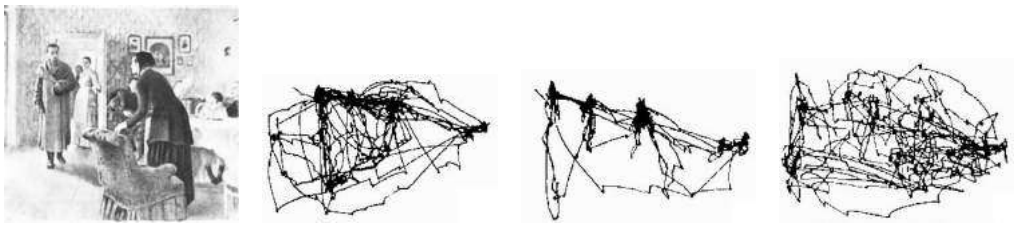


Figure 2.7: Guided Search Model [208, 203].

Top-down perceptual analysis involves cognitive factors, such as perceiver's expectations, memories, knowledge, prior experiences and tasks. These top-down signals can modulate the bottom-up process [232, 130, 74, 228]. For instance, Yarbus [217] illustrated in Figure 2.8 the top-down guidance of attention on a visual search task [210, 208, 211, 209].



(a) Image. (b) Free examination. (c) Give the age of people. (d) Remember the positions of objects.

Figure 2.8: Top-down guidance of attention during a visual search task [217, 228].

2.3.1 Computational Models

Itti et al. [96] followed the theory of Koch and Ullman [116] and as pioneers proposed the first implementation of a computational saliency model whose architecture is shown in Figure 2.9. For saliency computation they used three visual features – intensity, color based on opponent-color theory and orientation computed by Gabor filter[2]. The model utilizes

[2]Gabor filter, applied for instance in texture segmentation, can be viewed as a sinusoidal wave of a particular frequency and orientation multiplied by a Gaussian function [206].

center-surround operations to simulate the organization of ganglion retinal fields (see Section 2.1). An image is subsampled into a Gaussian pyramid and each level of the pyramid is represented by intensity, red, green, blue and yellow color channel and oriented edges in 4 different angles (0, 45, 90 and 135 degrees). A feature map is defined as a difference of two pyramid layers at different scales for intensity, red-green opponent, blue-yellow opponent and 4 orientations to detect local multi-scale feature contrasts that could cause the pop-out effect. The sum of feature maps of each visual feature is called a conspicuity map and their linear combinations lead to a saliency map. A sequence of human fixations are predicted using WTA and IOR. However, local irregularity may not be always salient since saliency may be caused by global irregularity and top-down aspects.

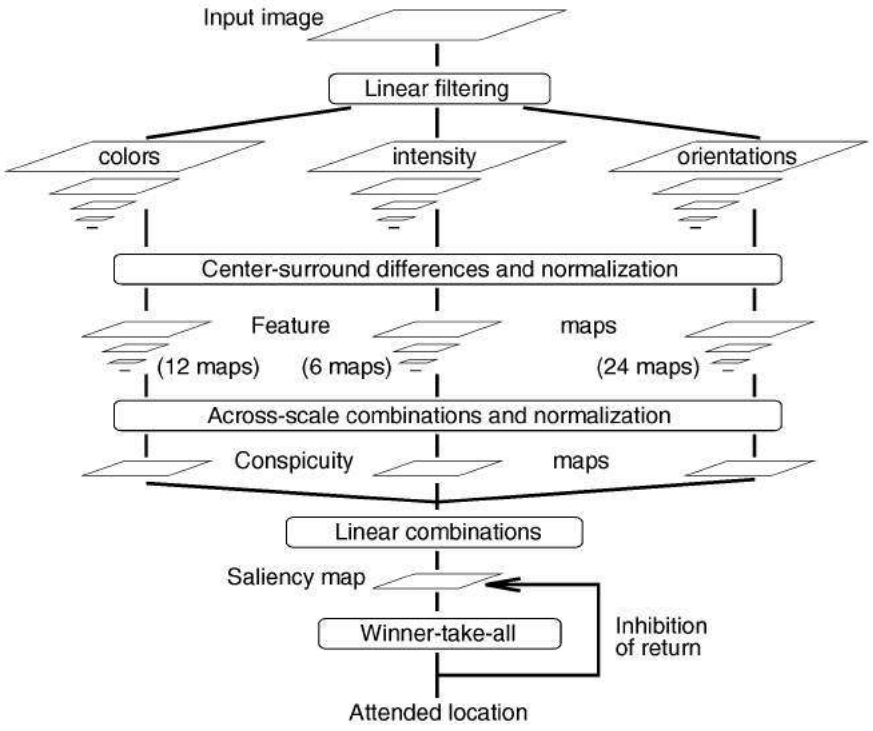


Figure 2.9: Architecture of a center-surround model by Itti et al. [96].

These center-surround filters have been adopted in many other computational models. For instance, Bruce and Tsotsos [24, 23] computes saliency using Shannon's self-information of high-level features derived by independent component analysis (ICA). Gao et al. [64] defined saliency as the Kullback-Leibler (KL) divergence between the features distributions of center and surround regions. Several other works integrate additional center-surround feature maps, including texture, flicker, motion, depth and isocentric curvedness [88, 94, 103, 198]. Harel et al. [80] measured similarity of Itti et al.'s features [96] in a fully connected graph built over all image locations to estimate saliency from global irregularity. Torralba et al. [194] and Zhang et al. [229] proposed Bayesian frameworks that combine bottom-up saliency that does not depend on the target with top-down knowledge of the target to estimate attention for visual search tasks. Target-independent bottom-up saliency of Torralba et al. [194] equals to $1/p(F|G)$, where F denotes local features at a given location and G represents global image features. Bottom-up saliency proposed by Zhang et al. [229] is based on self-information of center-surround features or learned ICA features. In contrast to center-surround saliency, Zhang and Sclaroff [226, 225] generated a set of binary (Boolean) maps by randomly thresholding color channels. Their model relies on the Gestalt principle

of figure-ground segragation. For each Boolean map, it identifies connected regions with closer outer contours which could belong to the foreground. Instead of saliency in the spatial domain, Hou and Zhang [86] computed the difference between the log amplitude spectrum and its smoothed version called the spectral residual. Hou et al. [85] extracted saliency in the frequency domain from the sign function of discrete cosine transform (DCT) coefficients, also known as the image signature.

The next group of computational models focuses on object-based attention. Some methods extract a single salient objects or multiple objects from location-based saliency maps [3, 138, 4]. Other works are purely region-based. They may employ object detectors and compute saliency directly at object level. Computational models often incorporate higher level semantic features such as persons, faces, cars, horizons and text [33, 105, 71, 152]. Judd et al. [105] learned the optimal weights of extracted low-level and high-level features from fixation data using the SVM classifier.

Since it is still difficult to achieve perfect object segmentation in complex scenes, several models split an image to regions and define saliency of each region. In general, regional saliency models based on global rarity define saliency of the region r_i comparing all other regions as $S_i = \sum_{j \neq i} w_{ij} D(r_i, r_j)$, where w_{ij} is a weight between regions r_i and r_j and $D(r_i, r_j)$ represents a contrast between the regions [19]. Cheng et al. [38] generated regions by a graph-based segmentation and described them by color histograms. Regional saliency depends on the region size, the color distance weighted by the color distribution and the spatial distance to all other regions in the image. Models proposed by Liu et al. [139] and Perazzi et al. [168] detected the global spatial and color contrasts too, but regions are represented by superpixels[3]. In addition to the superpixel color rarity, they measured the spatial distribution of superpixel colors. Colors with a low spatial variance are considered as the salient foreground. Li et al. [131] proposed a model based on mutual correlations of all superpixel pairs considering intensity, red-green and blue-yellow opponency. The model optimizes superpixel saliency to simultaneously meet three criteria – salient superpixels have a low correlation with other superpixels, superpixels in the image center are salient and correlated nearby superpixels are similarly salient. Jiang et al. [100] proposed local multi-scale superpixel contrasts to estimate saliency.

Recent research in computational saliency modelling achieves an impressive improvement of saliency estimation using deep learning network (DNN) architectures [32, 89, 40].

One of the first neural network techniques was introduced by Vig et al. [199]. They used features from layers of a deep model originally designed for face recognition. An optimal blend of features is learned using the SVM classifier. Fixation databases are relatively small in comparison with large image database for object recognition tasks. Recent DNN models with the highest prediction accuracy [122, 123, 40, 89] are therefore built on DNNs aimed for object or scene detection [185, 82, 120, 191] and then retrained on fixation data to estimate saliency [178].

[3]Superpixels are perceptually homogeneous areas with comparable size that should respect object boundaries [19]. For example, the Simple Linear Iterative Clustering (SLIC) algorithm [1] clusters pixels in a 5-dimensional space which consists of LAB color space and spatial coordinates.

2.4 Color Stimulus

Color vision discriminates differences in the wavelengths of light. The wavelength reflected from a stimulus determines its *hue*. The amplitude of waves determines the stimulus *brightness* and the purity of wavelengths determines the *saturation*. Color plays an important role in object detection and discrimination [74, 115].

There are two main theories of color perception.

The **trichromatic theory** proposed by Young [218] and Helmholtz [83] states that color perception is produced by three types of cone receptors with different, but overlapping spectral sensitivity (see Figure 2.2). They respond most to stimuli perceived as blue, yellow-green and orange-red, respectively. Perceived color is then the result of the relative stimulation of each cone receptor type [57, 108, 115, 75].

Hering [84] observed that some colors cannot coexist together, such as a reddish green and a bluish yellow. He also noted that the trichromatic theory does not explain negative *afterimages* which refer to the illusory perception of the complementary color to the one of a stimulus that has just been removed (see Figure 2.10) [115, 57].

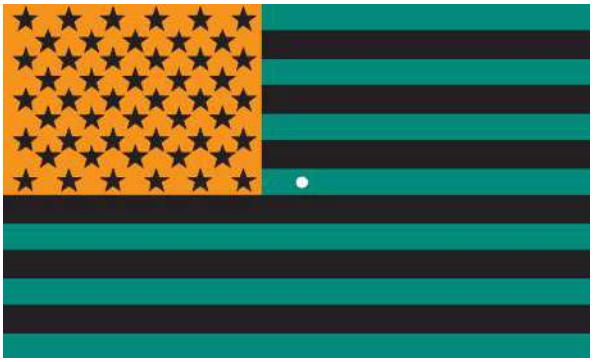


Figure 2.10: Fixating a central dot in the flag for about 30 seconds and then shifting the gaze to a white surface or blinking will cause the perception of a negative afterimage. The green area creates a red afterimage and the yellow area creates a blue afterimage [232].

Hence, Hering proposed **opponent-process theory** [84], according to which colors are processed in three complementary (opponent) pairs – *red-green*, *blue-yellow* and achromatic (*black-white*) mechanisms. Figure 2.11 illustrates that these mechanisms respond in opposite ways to different wavelengths or intensities. For instance, red-green mechanism responds positively to red and negatively to green [74, 115, 57, 108].

Later research [90] showed that both theories of color vision are correct. Signals from three cone receptors described in the trichromatic theory are sent to opponent retinal ganglion cells described in the opponent-process theory [115, 232, 57, 75].

In addition, color perception is affected by psychological aspects, such as culture and language of people and learned associations [155, 74, 51, 50]. Subjects of Gelasca et al.'s experiment [67] reported red which is associated with danger as the most salient color. Elliot et al. [51] showed that red has a negative effect on performance of activities. Lindsey at al. [136] found out that desaturated red targets are searched among distractors faster. They speculated that this effect arises from a perceptual specialization in human skin. Wool et al. [212] compared searching for targets of unique and non-unique hue. They did not observe

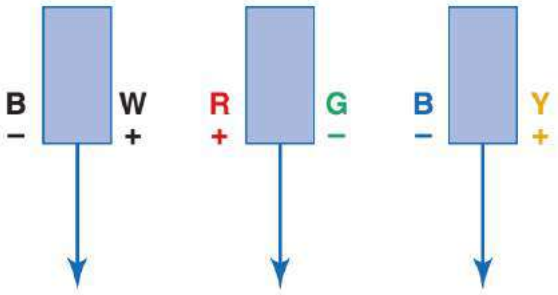


Figure 2.11: Opponent-process theory of color vision defines 3 complementary pairs – black-white (BW), red-green (RG) and blue-yellow (BY).

any differences, but they found that yellow targets were detected faster and with fewer saccades than blue targets. An experiment conducted by Etchebehere and Fedorovskaya [54] found mismatch between the numbers of fixated and reported stimuli. For instance, patches of red hue were highly reported, but less fixated than the rest of colors. On the other hand, green patches were highly fixated, but with a low probability of being reported.

2.4.1 Computational Models

Color saliency is implemented in most computational models. Bruce and Tsotsos [24] employed the RGB color space based on the trichromatic theory. Itti et al. [96] followed the opponent-process theory and computed color contrasts from red-green and blue-yellow channels. In addition to RGB, Borji and Itti [15] computed rarity in the Lab color space which approximates human color perception. Euclidean distance in Lab is used to estimate the perceived color difference [203]. Liu et al. [141] detected salient regions in the HSV color space. Duan et al. [48] used the YCbCr color space to estimate saliency. There is an implementation of Harel's et al. [80] saliency that employs the DKL [42] color space which models the cells with opponent-color character in the LGN[4]. Zhang et al. [227] added a color prior to saliency, so that warm colors are considered as more conspicuous.

2.5 Motion Stimulus

The most important function of attention to motion stimuli is linked to survival. Motion within the retinal image may occur when observers look at moving objects or move their eyes, the head or the entire body (*self-motion*) [13].

Motion processing is a complex process which includes motion detection of features in the retinal image, integration of feature movements to perceive coherent object motion and discrimination between object motion and self-motion. Motion perception also contributes to feature grouping, object segmentation, surface and depth perception [13, 75, 74]. Motion-produced information of the relative object distance is called *motion parallax*. As an observer

[4]The Derrington-Krauskopf-Lennie (DKL) color space [42] is represented by spherical coordinates. Let S, M and L be the response of three cone types. A luminance, vertical axis is defined as $L + M$. The chromatic plane is represented by $L - M$ and $S - (L + M)$ axes [121].

moves, objects closer to the observer move across the retina more rapidly than distant objects (see also Section 2.6).

Beside real object motion, we can also perceive illusory motion of stationary objects. The most important illusory motion is *apparent motion* which occurs, for example, when watching movies. Multiple stimuli at slightly different locations alternated over time may be perceived as a single moving object [75].

Gibson [69, 70, 68] proposed a theory that motion is perceived through variations in the retinal image, the so-called *optical flow*. Self-motion generates global optical flow, whereas relative motion of objects results in local optical flow in the retinal image. As an observer moves, all stimuli in the environment move in the opposite direction than the observer movement (see example in Figure 2.12). The place where the observe is heading is without the flow, called the focus from expansion (FOE). The flow magnitude gradually increases from the FOE, so that stimuli closest to the observer move across the retina most rapidly [75].

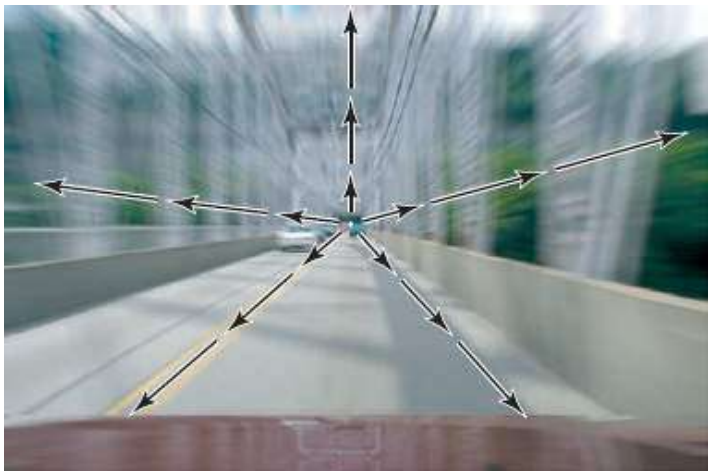


Figure 2.12: A moving car generates global optical flow in the retinal image [75].

Gregory [77] suggested that human visual system uses two different systems to distinguish object motion and self-induced motion. In addition to the retinal image whose change may arise from both types of movement, there is also a system that informs about eye and head movements [57].

Various experiments showed that human visual system is extremely sensitive to movements of living organisms, i.e. *biological motion* [74]. Human motion is perceived even with very limited information, e.g. from a sequence of point-light stimuli attached to joints of persons in a dark room [102, 101].

Behavioral studies have been also investigated the effect of flicker on visual perception. The sensitivity to flicker is limited. The increase of flicker frequency above a specific threshold leads to its appearance as a continuous light, which is referred to as the *flicker fusion* effect. The flicker fusion is important for video displays which appear continuous, despite of their flickering [74].

2.5.1 Computational Models

In contrast to spatial saliency, spatio-temporal saliency has been less explored. Computational models may define temporal saliency in video sequences either with bottom-up or

top-down strategy. Bottom-up models usually combine static and temporal features, e.g. motion, to detect spatio-temporal irregularity. On other hand, top-down models may learn prior information about images to capture unexpected, surprising stimuli or learn spatio-temporal features from an eye-tracking database of video sequences [130, 16, 93].

Itti et al. [94] extended the static center-surround model [96] by flicker and motion as dynamic features. Their saliency is estimated by comparing intensity channel and spatially shifted Gabor-based features over time and scale, respectively. In contrast to differentiating 2D features between subsequent video frames, Rapantzikos et al. [174] generated spatio-temporal cubes of intensity, color, 2D and 3D orientations from a video sequence. Feature cubes are defined at multiple scales by 3D Gaussian filters and across-scale differences are fused into a saliency volume. Zhai and Shah [221] estimated multiple homographies between consecutive video frames by applying the Random sample consensus (RANSAC) algorithm[5]. Projection errors of points define temporal saliency. Cui et al. [41] presented a spatio-temporal model in the spectral domain. A spectral residual approach is performed on horizontal and vertical temporal slices across video sequences. Loy et al. [144] defined the spectral residual on the optical flow phase and magnitude fields. Liu et al. [140] enhanced a superpixel-based approach [139] by motion field calculated by an optical flow estimation algorithm. The proposed model computes the contrast between the motion histogram of each superpixel and the motion histogram of whole image. In addition, it measures the color distance between each superpixel and its projection in a consecutive frame. Marat et al. [149] proposed a spatio-temporal model inspired by the parvocellular-like and the magnocellular-like retinal output for static and temporal saliency, respectively. Fang et al. [58] defined spatio-temporal saliency from the DCT coefficients and motion vectors for each block of MPEG-4 video[6]. A center-surround spatio-temporal model by Mahadevan and Vasconcelos [146] is based on the KL divergence between dynamic texture parameters.

A specific category of spatio-temporal saliency is egocentric saliency. Since egocentric videos are recorded from the first-person perspective, these computational models should incorporate motion induced by head and eye movements in contrast to traditional models [215]. Yamada et al. [214] estimated camera's rotation velocity and direction of movement. A work proposed by Li et al. [132] learns gaze in egocentric videos from head and hand cues using a random regression forest.

Itti and Baldi [92] defined saliency from surprise which is a significant change of observer's beliefs. It is modelled in a Bayesian framework by the KL divergence between posterior and prior distributions over space and time.

2.6 Depth Stimulus

Depth perception reconstructs 2D retinal image to perceive objects three-dimensionally. Beside real objects in the environment, depth can be also perceived from 2D images [115, 74].

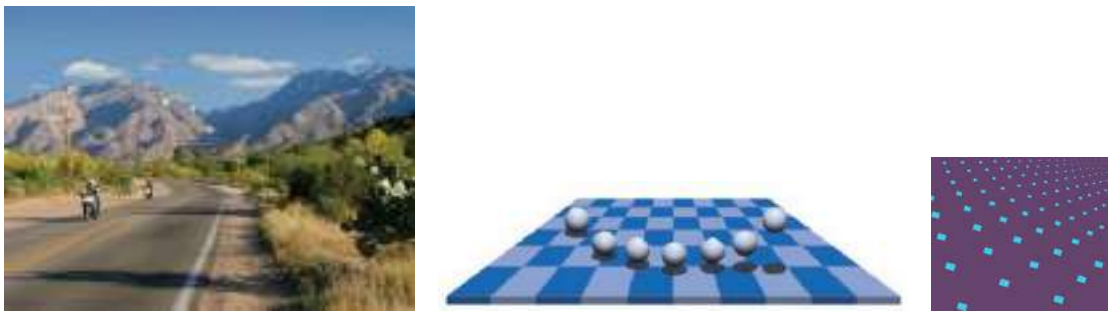
There are three types of cues to perceive depth [75, 57]:

[5]RANSAC computes iteratively a homography matrix from a random subset of points and checks whether remaining points are consistent with this estimation (inliers) or not (outliers). The algorithm keeps the homography with the lowest number of outliers [188].

[6]MPEG-4 ASP standard processes video frames in 16×16 units called *macroblocks*. Each macroblock consists of four luminance, one blue chrome and one red chrome 8×8 blocks encoded using the DCT transform. In addition, motion vectors are estimated at macroblock level [58].

1. **Oculomotor** cues depend on sensations of eye movements and eye muscles.
2. **Monocular** cues require the retinal image only from one eye.
3. **Binocular** cues depend on the combination of retinal images from both eyes.

Pictorial cues depicted in images enable depth perception in monocular vision (see examples in Figure 2.13). For instance, *occlusion* is a monocular depth cue which occurs when an object partially hidden by another objects is seen as being more distant. Depth information is also provided by *relative height* of objects. Objects below the horizon that are positioned relatively higher in a picture are seen as being farther away. The opposite process implies for objects above the horizon. A further cue is *relative size* of objects. When there are equally sized objects, the more distant object is relatively of lower size. *Perspective convergence* provides another monocular cue to depth. Parallel lines pointing directly away are perceived as converging. The cue to depth is also *familiar size* of objects known from experience about the standard object size. According to the cue of *atmospheric perspective*, objects that are farther away appear less sharp. Another useful cue is *texture gradient*. Texture is perceived as denser and finer as distance from an observer increases. *Shadows* of objects provide additional information about their relative distance to an observer [75, 13, 115, 57].



(a) Occlusion, relative height and size, perspective convergence.

(b) Shadows.

(c) Texture.

Figure 2.13: Examples of pictorial cues to depth. The cactus occludes the farther hill, the farther motorcycle is relatively higher and smaller than the nearer one and the sides of the motorway converge in the distance (left). Shadows of the balls determine their relative position to the chessboard (middle). Texture appears to be denser as distance increases (right) [75, 115].

Beside static cues, there are also **motion-produced** monocular **cues**. The effect known as *motion parallax* occurs when an observer moves. Objects closer to the observer appear to move faster than objects at the far distance. *Deletion and accretion* are cues resulting from an observer's sideways motion when parts of overlapping surfaces become covered or uncovered [75].

Stereopsis is depth perception based on slight differences between left and right retinal images, also known as *binocular disparity*. The combination of both retinal images and the retinal disparity detection lead to perception of the environment in 3D [57, 115].

Lang et al. [125] conducted an eye-tracking experiment with depth cues from 2D and 3D images. Comparing fixation distributions of both version, the authors concluded that depth cues modulate visual attention more significantly at farther distances. Furthermore, locations at a closer distance were fixated more frequently. They also revealed a non-linear relationship between saliency and object distance. According to the study of Jansen et al. [98], closer

locations in both 2D and 3D images are fixated earlier. They did not found a significant effect of binocular disparity on 2D saliency. The experiment of Wang et al. [202] employed stereoscopic images in which varying number of identical circular stimuli organized in a circle were displayed at different depth positions. Their research found most fixations at the object closest to an observer too. However, the results showed a slight decrease of the average depth position of fixations over time. Designh et al.'s study [43] used 2D images with 10 quantized levels of depth range. They found out that objects located at the 4^{th} depth level got most fixations, mainly of longer duration. In contrast, closest objects were primarily fixated for a shorter period of time. Ramasamy et al. [173] analyzed gaze data for 2D and 3D video clips. A sequence with a long deep hallway revealed a different gaze behavior between both video versions. Participants' gaze was highly focused at furthest locations for a stereoscopic version of the clip. On the other hand, fixations were more spread out in the 2D version. Experiments comparing monocular and binocular vision in real environments showed that stereopsis enhances depth discrimination and estimation even at larger distances [5, 164].

2.6.1 Computational Models

A majority of computational models focus on 2D saliency. However, Itti and Koch [95] suggested to include stereo disparity in a bottom-up model of attention. Depth information estimated from a depth camera, e.g. Kinect, may be used as a weighting factor of 2D saliency maps or as an additional saliency map fused with 2D saliency. Besides, several models estimate attention from stereo images [201].

Some computational models compute 2D saliency map separately and weight the saliency by depth so that the highest value is assigned to objects closest to an observer [147, 34, 231]. Gautier and Le Meur [66] used depth information to separate background and foreground. Ouerhani and Hugli [163] applied center-surround differences by Itti et al. [96] to compute an additional depth conspicuity map. Ju et al. [103] adopted center-surround operations as depth difference in 8 directions. Several works defined saliency as global region-based contrast of depth channel [160, 43, 224]. Lang et al. [125] employed a Gaussian mixture model to learn depth saliency. Wang et al. [201] computed depth saliency using a Bayesian approach as $p(C = 1|F)$, where C indicates whether a given location is fixated or not and F equals to the center-surround depth contrast.

2.7 Shape Stimulus

In contrast to intensity, color and orientation, shape is a complex visual feature combining boundary and surface properties of a particular object. Shape perception is an important part of object recognition. It is a complex process integrating many visual features such as contours, color, texture, shading and depth. Object perception includes figure-ground separation, representation of object properties and association with learned objects [13].

Human visual system is very sensitive to regions of high curvature, such as angles. According to the theory of structural representation, objects are represented by common geometrical shape of their components. The representation is a result of neurons tuned for various position and shapes. To perceive shape of object components, neurons are tuned for *geometric*

derivatives such as contour orientation (first-order), contour curvature (second-order), contour spirality (third-order) and surface curvature [44, 13, 142].

However, shape saliency is affected not only by low-level features, but also by memory representations of objects. More complex shapes may be processed in specialized parts of visual system. For instance, humans are very sensitive to shapes resembling *human faces*. In addition, attention can be also influenced by the category of objects and their mutual relationships [13].

Visual shape perception is also described by *Gestalt principles* [117, 207, 118]. For example, there is a perceptual grouping of similar and neighbor objects and incomplete object parts (see Section 2.3).

Texton Theory [106] states that human visual system detects a group of features called texton. Textons are characterized by elongated blobs with specific visual properties, terminators and crossings of line segments.

2.7.1 Computational Models

Standard computational models usually cover shape saliency only partially. For example, Itti et al.'s [96] center-surround contrasts participate in shape perception, but their model does not involve global shape properties and differences.

This model is enhanced by saliency of object shapes in a work presented by Chen et al. [37]. In order to compute shape saliency, they compared shapes of each object pair aligned by their centroids using the Jaccard similarity index. Shape saliency proposed by Valenti et al. [198] is based on the curvedness defined as $\sqrt{L_{xx}^2 + 2L_{xy} + L_{yy}^2}$ (L stands for luminance) and clustering using isophote information[7]. Chen and Zhang [36] employed shape information to detect salient regions in satellite images. Since shape contrast is defined as a ratio of the square root of region area to its perimeter, lower saliency is assigned to roads in images. Object-based saliency is included in an approach by Jiang et al. [100]. Salient objects differ from the neighborhood and their boundary is closed. Qi et al. [172] utilized training of Restricted Boltzmann Machine and Conditional Random Field (CRF) to obtain global shape saliency and local spatial saliency. A work presented by Li et al. [133] predicts saliency using a random forest. Their model is trained on simple global shape descriptions of segmented regions including area, centroid, eccentricity, perimeter and human fixations. Following the figure-ground segregation proposed by Gestaltists, Zhang and Sclaroff [226, 225] detected figures as regions with closer outer contours and the background as regions connected to image borders. A method introduced by Kim and Pavlovic [112] trains a convolutional neural network on image patches to predict shape classes and saliency.

Shape saliency is also applied for 3D objects. The model proposed by Lara et al. [126] compares objects with their bounding rectangles to assign lower saliency values to flatter surface. Song et al. [187] detected salient regions by a CRF framework for 3D meshes.

[7] Isophotes are curves connecting points of equal brightness [198].

2.8 Emotional Factors

In recent years, several studies confirmed the connections between emotional stimuli, current emotional state, visual perception and attention. The perception of emotional stimuli predominates over emotionally neutral stimuli [57, 74, 170, 8, 176, 220]. There are various types of emotions which differ in universal facial expressions, including anger, disgust, fear, happiness, sadness and surprise [49].

Emotional experience involves both stimulus-driven and knowledge-based processes [57]. Compton [39] defined a two-level process in the processing of emotional information. First, the importance of an emotionally tuned stimulus is captured in the amygdala. In the next step, emotionally-driven stimuli are more successful in reaching the selective attention. This short process includes bottom-up inputs as well as top-down processing.

When searching for targets in a rapid stream of stimuli, observers often detect the first target, but fail to detect the second target which is displayed shortly after the first one. This phenomenon, also known as the *attention blink*, is reduced when a highly emotional stimuli is used for the second target [177, 74].

Multiple studies aimed at the broadening effect of positive emotions on visual attention [63, 200]. Fredrickson's *broaden-and-build theory* [62] is the most widespread theory about the effects of emotions on attention. The theory states that positive emotions, such as joy, interest, contentment and love widen individual's attention and thinking. According to the theory, positive emotions create urges to act in a specific way, for instance joy invokes the urge to play.

Several experimental studies showed that emotions enhance attention on the performance and *perceived contrast* [170, 8]. Killgore and Yurgelun-Todd [111] suggested that positive mood may affect the early stage of sensory processing. The work of Grol et al. [78] went further and specified a relationship between positive mood and broadened attention. They found out, that positive emotion was associated only with broadening attention for self-related stimuli and pointed to the target of attention as the key aspect. Eye-tracking experiments of Wadlinger and Isaacowitz [200] showed that visual attention in positive mood is connected explicitly with highly positive target stimuli. The experiment conducted by Talarico et al. [192] showed that positive emotions enhance recall of autobiographical memories with emphasis on peripheral details. Ohman et al. [161] studied attention through fear-relevant and fear-irrelevant pictures. In a search task, fear-relevant stimuli were found much faster regardless of the number of distractors and the location of a target stimulus. A later research revealed a lack of attention control caused by the negative valence of stimuli [157]. According to Kaspar et al. [110], the induction of positive emotions increases the attention and the memory for negative stimuli. Participants were induced to positive or negative mood, then they were asked to search and read some webpages with online news. After a memory test, they found a relationship between the attentional preference and the memory preference for negative news only in a positive mood condition. Negative discrete emotions are also significant in the process of decision making. A study of Ferrer et al. [60] revealed a decreasing effect of accurate perceptions caused by negative emotions. Rowe et al. [180] confirmed the assumptions of broaden theory and claim that people in positive mood produced more semantic associations in contrast to people in sad or neutral mood. Maekawa et al. [145] studied how happiness influences visual search times. While there was no effect in a search for a single open circle among identical closed circles, happiness leads to an efficient search

for a single closed circle among differently oriented open circle.

Negatively and positively arousing stimuli may also act as visual attention distractors, the effect also known as *emotion-induction blindness*. The experiment by Most et al. [158] showed that positive and negative affective stimuli caused spontaneous attentional blink while performing the task. Stimuli with positive arousal represented by erotic pictures involuntary captures respondents' attention. Pecher et al. [166] found out that listening to happy music while driving may distract drivers' attention. When happy music was playing in a car, drivers' mean speed was decreased and their peripheral control has been worsened.

Visual search performance could be also affected by *engagement* that relates with mood. Smilek et al.'s research [186] found benefits of lower engagement in a visual search task. Instructing observers to passively let a target be seen may be associated with automatic visual processing and therefore lead to more efficient search than actively directing attention to the target. Similarly, Olivers and Nieuwenhuis [162] showed that a slight disengagement from a current task reduces the attentional blink. Jefferies et al. [99] induced observers into different moods (positive, negative) as well as different engagement levels (high, low). Sad observers with low arousal were best at avoiding the attentional blink.

2.8.1 Computational Models

Only few attempts have been made to include emotional factors in a saliency model. These computational models estimate saliency for affective scenes. Affective analysis employs facial expression estimation, affective object detection (e.g. snakes, worms and flowers), emotional reactions to colors or emotion classifiers trained on images labeled with emotional responses [137, 46]. Instead of saliency computation, some works [167, 190] employ a convolutional network to predict emotional stimuli.

2.9 Visual Attention in Visualizations

Memory, visual attention and perception play a critical role in the design of visualizations. Visualization designers use a large variety of visual channels to effectively encode data. In addition to bottom-up factors of attention, top-down factors are incorporated in scene perception when users interpret visualizations. Visual search is an important component of the process of interpreting visualizations. It is the process of finding a specific target object in a scene among non-targets. Visual attention thereby guides the user's gaze and the visual search, respectively. Understanding visual attention is therefore essential for selecting appropriate visual channels and designing effective visualizations.

Human gaze behavior is influenced by activities in visualizations. Data analysis using visualizations can be divided into three categories [155]:

1. *exploratory analysis*: to formulate a new hypothesis about the data,
2. *confirmatory analysis*: to confirm or reject given hypotheses about the data,
3. *presentation*: to communicate facts efficiently and effectively.

2.9.1 Computational Models

Standard saliency models designed for natural images have been also used in visualization research to predict attention and derive quality measures, respectively [65, 97, 10, 179].

While the prediction ability of these models is quite accurate for simple stimuli and natural images [14, 165, 169], it has been shown that these models' accuracy for predicting visual attention in visualizations is significantly poorer [79]. There are some notable differences between natural images and classic charts used in information visualization. Graphical marks, such as dots or lines, are usually abstract and simple compared to complex objects in natural images. Also, the background is mostly uniform and the visualizations contain a lot of textual information, such as labels and legends. Graphical marks and visual channels are chosen by a visualization designer according to design guidelines and visualization domain knowledge with the goal to expressively and effectively represent the underlying data. Thereby, visualization designers choose their visual channels to maximize the amount of information displayed [81]. Matzen et al. [152] also noted that most saliency models tend to omit fine-grained visual features, like thin lines, which are highly relevant for information visualization.

Therefore, specialized saliency models have been developed. Lee et al. [128], for instance, introduced a saliency model for categorical map visualizations. They defined point saliency as color difference of each point against its neighborhood. The class visibility quantifies the point saliency values that correspond to a given category. Kim and Varshney [114] proposed center-surround operations on voxels to guide attention to selected regions in volume visualizations. Matzen et al. [152] proposed a novel saliency model tailored towards information visualization that combines the model of Itti et al. [96] with text saliency and could thereby increase the performance of the saliency model significantly.

2.10 Evaluation Metrics of Computational Models

There are various metrics how to compare computed saliency maps and human fixations. They can be divided into two groups – *location-based* and *distribution-based*, which evaluate saliency maps against discrete fixations or continuous fixation heatmaps, respectively [31, 18, 127, 130, 16].

2.10.1 Location-Based Metrics

The most widely used evaluation metric is **the Area under the Receiver Operating Characteristic Curve (AUC)**. The Receiver Operating Characteristic (ROC) curve represents the trade-off between the true positive (TP) rate and the false positive (FP) rate. A saliency map is treated as a binary classifier. Saliency pixels at fixations and some non-fixated pixels are extracted. Fixations with saliency above a threshold that is gradually increasing and non-fixations above the threshold are considered as TPs and FPs, respectively. Then, the ROC is plotted and the area under the curve (AUC) is computed. An AUC value of 1 corresponds to a perfect fit between fixation map and saliency map, while 0.5 corresponds to chance level [31, 225].

To report this score for an image dataset, AUC is usually computed individually for each image and the average is computed[130][8].

There are various definitions of FPs. Some variants of AUC define FPs as all non-fixated locations (e.g. *AUC-Judd* [105]), others use only a uniform random sample of non-fixations, mostly of the same size as fixation points (e.g. *AUC-Borji* [17]).

Because of a center bias[9], a central Gaussian distribution would mostly achieve a high AUC score [104]. To tackle the bias, the *shuffled AUC* score [18] samples FPs from fixations from other images. Since some FPs are centrally positioned fixations, i.e. locations fixated independent of the content, this score discounts the fixations affected by the center bias [31].

High AUC scores are calculated for saliency maps with higher values at fixations, whereas non-fixated low saliency values are mostly ignored [31]. Judd et al. [104] showed that blurring the map increases the AUC scores for most computational models.

As Bylinskii et al. [32] noted, the AUC metrics have almost saturated especially thanks to neural network models. The saliency evaluation should therefore move towards other metrics which are still able to differentiate the models, for instance the **Normalized Scanpath Saliency (NSS)**. The score equals to the average saliency at fixation locations in a saliency map normalized to have a zero mean and a unit standard deviation. For the NSS, a value of 0 corresponds to chance level, and the higher the NSS score, the better the fit. In contrast to the AUC metric, the NSS score is affected by all FPs. Therefore, it could be considered as the the fairest location-based evaluation metric [31].

2.10.2 Distribution-Based Metrics

In contrast to location-based metrics, distribution-based metrics employ blurred fixations, mainly using a Gaussian filter [31].

Saliency S and continuous fixation maps F can be compared using the **Correlation Coefficient (CC)**[10], defined as $\frac{cov(S,F)}{\sigma(S)\times\sigma(F)}$. The perfect linear relationship is represented by the maximum (1) or minimum score (-1). The CC score is considered as the fairest distribution-based metric because it is equally affected by false positives and negatives, as the NSS score [130, 31, 179].

The **similarity metric (SIM)** is defined as $\sum_x \min(S(x), F(x))$, where $\sum_x S(x) = 1$ and $\sum_x F(x) = 1$. The SIM score is therefore more sensitive to false negatives than false positives. The score of 1 represents the equal distributions of both maps [179, 31].

The **Kullback-Leibler divergence (KL-div)** measures the dissimilarity between two distributions as follows: $\sum_x F(x) \log(\frac{F(x)}{S(x)+\epsilon} + \epsilon)$, where ϵ is a small constant, $S(x) = \frac{S(x)}{\sum_i S(i)+\epsilon}$ and $F(x) = \frac{F(x)}{\sum_i F(i)+\epsilon}$. For the KL-div, a value of 0 represents the equal distributions. and the lower the KL-div score is, the better it fits the fixation data. As the SIM score, the KL-div penalizes more false negatives than positives [179, 31].

[8] An alternative method is to compute only a single AUC value using TPs and FPs from all images [130].

[9] Attention is biased towards the center of an image, i.e. a center bias [31].

[10] Most evaluations use the Pearson's r.

2.10.3 Human Inter-Observer

Even though some of the described metrics are bounded, their perfect score may not be reachable. Therefore, some researches compute the score of the **human inter-observer (IO)**. It generates output maps from fixations of all participants except one under the test. The score represents the fixation consistency across users and can be thereby considered as upper bound to the evaluation scores [225, 89, 31].

Chapter 3

Contribution

This chapter presents contributions of this thesis structured by the factors that affect human visual attention:

- static stimuli such as color, shape and depth,
- dynamic stimuli such as motion,
- observer's emotions,
- task-based analysis of visualizations.

These stimulus-driven and goal-directed factors are investigated through novel computational saliency models[1] and eye-tracking experiments with still images and egocentric videos. The outputs of our studies could help to understand human gaze behavior and improve saliency estimation.

3.1 Egocentric Motion Saliency Modelling

Since egocentric saliency has not been widely explored so far, we introduced own superpixel-based saliency model for egocentric videos in the author's master thesis. This section continues in analysis of this model. In contrast to the master thesis, we evaluated this model on a larger dataset and compared with other existing saliency models. Evaluation has been done on a natural shopping task recorded by eye-tracking glasses. This model has been already published in [234] and [237].

3.1.1 Motivation

Estimation of egocentric gaze behavior in a natural environment has gained increased interest with the advent of miniaturized wearable cameras, such as GoPro and Google Glass. A person is taking visual measurements about the world in a sequence of fixations which contain relevant information about the most salient parts of the environment and the goals of the actor. Prediction of gaze from the first-person perspective becomes increasingly relevant in order to interpret the continuous video stream in daily activities and deduce appropriate

[1] We used NSS, AUC-Borji and SIM metrics for saliency evaluation (see Section 2.10).

analytics and recommendations in the domains of health, social interaction analysis, traffic security, or in market research.

Recent research on the first-person vision and egocentric video analysis [12] employed salient features for the purpose of activity recognition [214]. In the context of hand-eye coordination it has been exploited for video annotation that the distribution of both visual features and object occurrences in the vicinity of the gaze point is correlated with the verb-object pair describing the action [59]. Implicit cues from visual features, such as hand location and pose, head and hand motion are useful features in this context [132]. Even in general settings on gaze prediction without significant focus on hand-eye coordination, camera motion estimation has been approved to represent a strong cue for gaze prediction [151]. Gaze provides the means to personalize the summary of a video sequence and provide a relevant feature for combinatorial optimization [213].

3.1.2 Analyzed Computational Model

First, we briefly summarize the saliency model proposed in the master thesis [234, 237]. Our model uses a superpixel segmentation [1] to at least partially implement object-based attention. Each superpixel is described by static (intensity, color and orientation) and dynamic (motion) features in multiple scales. Since human gaze is also directed to unexpected, surprising stimuli, saliency estimation includes motion surprise too.

Each video frame is decomposed into intensity, red, green, blue and yellow colors and orientation of gradients by applying Sobel filter. Motion between consecutive frames is calculated by an optical flow algorithm, as shown in Figure 3.1 [35][2]. Estimated motion field is characterized by its magnitude and orientation.

The distribution of each feature within a superpixel is represented by a histogram, so that we calculate 8 histograms for each superpixel (6 histograms for static features, 2 histograms for temporal features).

(a) Video frames. (b) Optical flow [35]. (c) Motion field.

Figure 3.1: Motion estimation (flow orientation and magnitude are encoded by hue and saturation channels of motion field, respectively).

[2]We used the optical flow introduced by Chambolle and Pock [35] which improves the TV-L1 optical flow proposed by Zach et al. [219] to overcome original assumption that intensity remains constant over time. Let Ω be the image plane, $v = (v_1, v_2)^T : \Omega \rightarrow \mathbb{R}^2$ be the motion field and $u : \Omega \rightarrow \mathbb{R}$ model the varying illumination. The algorithm solves $\min_{u,v} \int_\Omega |Du| \int_\Omega |Dv| + \lambda \|\rho(u, v)\|_1$, where $\rho(u, v) = I_t + (\nabla I)^T (v - v^0) + \beta u$, I_t is the time derivative, ∇I is the spatial image gradient, v^0 is some given motion field and symbols λ and β represent weights.

To follow the multi-scale approach of Itti et al. [96], we subsample superpixel boundaries into a pyramid and generate superpixel representations for each layer. The layer $k+1$ is produces as follows:

1. downsample a map with superpixel boundaries,
2. compute a histogram of each superpixel using the layer k.

Saliency from spatio-temporal contrasts is analogous to Itti et al's model [96]. For each pixel of a frame, we identify corresponding superpixels in a center layer ($c = \{0, 1, 2\}$) and a surround layer ($s = c + \delta$; $\delta = \{0, 1, 2\}$) and compare their histograms.

Saliency from spatial contrast S_{SC} is computed from the intensity, red-green opponency, blue-yellow opponency and orientation feature maps using the correlation coefficient or the mean feature value at superpixel level.

Saliency from temporal contrast S_{TC} is based on the average flow vectors $\mathbf{v} = [\varphi, r]$ within each superpixel, where φ is orientation and r is magnitude. The center-surround difference of flow vectors highlights local disturbances of retinal optical field which may arise from object motions:

$$M(x, y) = \|\mathbf{v}_s(x, y) - \mathbf{v}_c(x, y)\| \tag{3.1}$$

Since attention in egocentric vision is affected by prior knowledge, observer's gaze is also directed towards unexpected, surprising object movements. To estimate motion surprise at a given location, we compare prior knowledge about motion field with the actual frame for each pixel:

$$S_{MS} = \|\mathbf{v}_{MEM_t}(x, y) - \mathbf{v}_t(x, y)\| \tag{3.2}$$

Prior knowledge about the flow is denoted by MEM_t, where t is time. This knowledge is updated with each frame as: $MEM_{t+1}(x, y) = (1 - \eta)MEM_t(x, y) + \eta v_t(x, y)$, where $\eta = 0.05$.

3.1.3 Evaluation

The dataset (2 videos, 860 frames in total of 1280×960 size) was recorded using eye-tracking glasses at a shopping mall where participants were asked to find specific products.

Our temporal saliency is based on the Gibson's approach to perception [69, 70, 68]. We assume that flow vectors different from their surround are generated by object movements and are therefore salient (see Figure 3.2). Moreover, saliency from temporal surprise may suppress regions with constant motion at a particular location, e.g. a trolley in Figure 3.3.

The performance of our model has been compared with the spatial location-based model by Itti et al. [96] and the spatio-temporal superpixel-based model by Liu et al. [140] (see example in Figure 3.4). Our model combines three types of saliency as $S = (1 - \lambda)S_{SC} + \frac{\lambda}{2}S_{TC} + \frac{\lambda}{2}S_{MS}$. In contrast to our saliency from local contrasts and motion surprise, Liu et al. [140] defined saliency from global spatial contrasts, motion distinctiveness to global frame-level motion and color similarity of corresponding superpixels over time.

We used two evaluation scores – AUC and NSS (see definitions of evaluation scores in Section 2.10). To compute AUC, saliency maps have been smoothed with a Gaussian kernel with standard deviations σ from 0.01 to 0.13 in image width (in steps of 0.01) and the maximum AUC have been taken.

(a) Video frame. (b) Temporal saliency.

Figure 3.2: Temporal saliency selected human and trolley movements as regions that would attract attention.

(a) Video frame. (b) Motion surprise.

Figure 3.3: Motion surprise suppresses a trolley with constant motion.

(a) Video frame. (b) Our saliency.

(c) Itti et al. [96]. (d) Liu et al. [140].

Figure 3.4: Evaluated saliency maps. Location of participant's fixation is labelled with a red circle in the frame.

We report both results in Table 3.1 and 3.2. We set λ to 0.4 which leads to the highest prediction accuracy for this dataset. This means that static saliency predominates over motion saliency despite of multiple moving objects in participants' views.

Table 3.1: Individual AUC scores.

Participant (video)	Our	Itti et al. [96]	Liu et al. [140]
#1	**.661**	.595	.649
#2	.749	**.756**	.742

Table 3.2: Individual NSS scores.

Participant (video)	Our	Itti et al. [96]	Liu et al. [140]
#1	**0.59**	0.35	**0.59**
#2	0.87	0.91	**0.94**

Even though we reduce temporal saliency only on novel movements, varying performance of computational models indicates that static and temporal saliency do not affect participants' attention equally. We therefore suggest to combine feature saliency maps dynamically, e.g. learn the effect size of each feature on visual attention from egocentric fixation datasets. In addition, a model should use additional attention aspects of egocentric vision including depth information and top-down features of attention, e.g. detection of biological motion. Saliency from surprise could also incorporate static features to model prior knowledge and detect unexpected locations with higher precision.

3.1.4 Summary

In this section, we have evaluated the egocentric spatio-temporal model proposed in the master thesis. The method fuses spatial, motion and surprise-based factors of attention. Fixation data from the viewer's camera indicate that spatial contrasts dominates over motion contrasts, though both saliency types have a substantial impact on egocentric vision. Furthermore, the complex scenes could affected by additional aspects of attention, such as prior knowledge and object detection, that should be included in saliency modelling, in future.

3.2 Visual Attention to Color

Color is the fundamental component of visual attention. Saliency is usually associated with color contrasts. Beside this bottom-up perspective, some recent works indicate that psychological aspects should be considered too [50]. However, relatively little research has been done on potential impacts of color psychology on attention. To our best knowledge, a publicly available fixation dataset specialized on color feature does not exist. We therefore conducted a novel eye-tracking experiment with color stimuli and made it publicly available. We studied whether color differences can reliably model color saliency or particular colors are preferably fixated regardless of scene content, i.e. color prior.

3.2.1 Motivation

In terms of color saliency, attention may not be affected only by rare and contrast regions. Recent research suggests to involve high-level psychological aspects in attention modelling too [50]. However, the effect of color psychology on selective attention has been little studied so far. Recent studies assume that particular colors have a considerable higher fixation probability, for example warm colors as red and yellow linked with danger and warning [54, 67, 136, 212, 227]. In addition, colors could affect performance on tasks, alertness or emotions [50]. This study could help to contribute to this complex research area and enhance computational models by color prior.

3.2.2 Experimental Study on Color Saliency

Eye-tracking experiments which simultaneously explore the relationships between color psychology, color contrast saliency and selective attention are lacking. Our experimental design is inspired by Gelasca et al.'s work [67] which employed scenes with differently homogeneously colored objects. In contrast to their study which asked subjects to report the most attractive colors, we analyzed color saliency on fixation data. We investigated the effect of color contrasts and potential color preferences. Data from our experiment are publicly available[3].

Hypothesis

Colors associated with warning and danger are highly attentive. We reason that a primary role of attention linked to survival results in a higher number of fixations of red and yellow colors. We therefore expect that the learned signals to potentially dangerous situations (top-down attention) reduce the bottom-up effect of color saliency – color contrasts.

Image Data

Natural scenes contain many stimuli that attract attention. Therefore, we used own simple synthetic images which contain homogeneously colored squares or circles. Our image dataset consists of 75 scenes with an uniform background. Each scene consists of 3 up to 9 objects that are either cluttered or uniformly distributed (see Figure 3.5). Some scenes contains objects of the same color. We employed 8 object colors that are defined in Table 3.3.

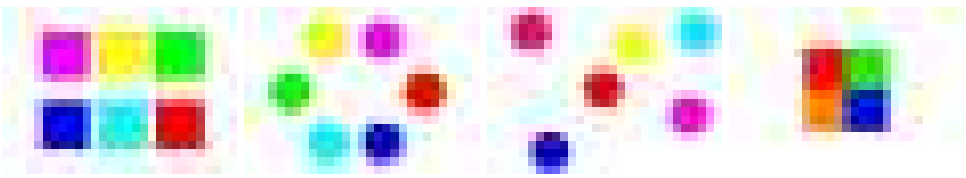


Figure 3.5: Examples scenes used in the experiment.

[3] https://vgg.fiit.stuba.sk/2018-08/color-saliency/

Table 3.3: Colors used in the experiment.

Color name	CIELAB	# objects
red	53.2, 80.1, 67.2	55
green	87.7, -86.2, 83.2	48
blue	32.3, 79.2, -107.9	58
yellow	97.1, -21.55, 94.5	54
cyan	91.1, -48.07, -14.1	36
magenta	60.31, 98.25, 60.8	41
pink	54.9, 84.55, 4.06	29
orange	67.05, 42.8, 74	29

Participants and Apparatus

15 students and members of academical staff voluntary participated in our experiment (12 males, 3 females) whose age ranges from 21 to 60 years. All participants gave their informed consent to the study and received an explanation of the experiment.

We record participants' gaze using Tobii X2-60 eye-trackers at 60 Hz and processed by Tobii I-VT fixation filter. Images were displayed on 24.1-inch monitors with a resolution of 1920 $\times$ 1080 pixels at viewing distance of approximately 60 cm.

Experimental Design and Procedure

Participants were shown 30 images from our dataset, each for 2 seconds. The image order was counterbalanced with a Latin square across participants. Each image was followed by a random cluttered image to reset participants' attention.

Measures and Analysis

Since we investigated the early attention stage, we took only first three fixations on objects into account. Then, we computed a ratio of fixations for each color in all scenes where a given color was used at least for one object.

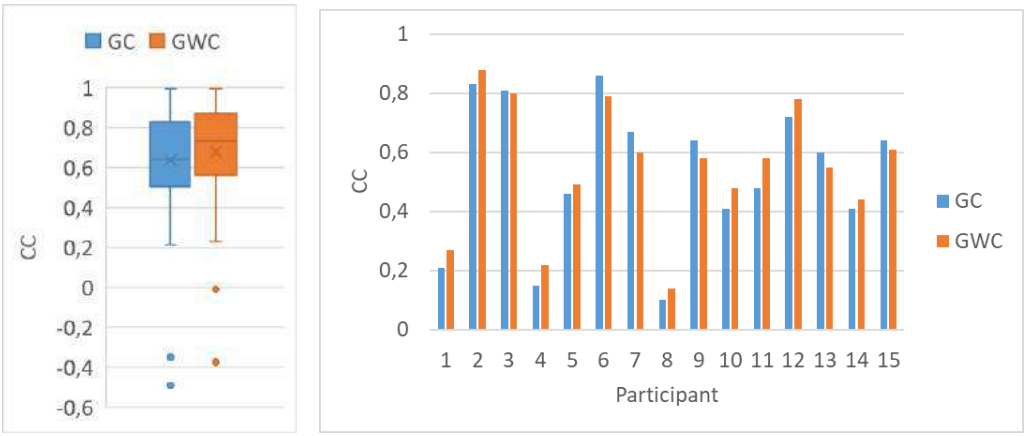
To analyze color contrasts, we used the CIELAB color space because it models human color perception. We defined the *global contrast* as $GC(O_i) = \frac{1}{N-1}\sum_j^N D_c(O_i, O_j)$ and the *spatially weighted global contrast* as $GWC(O_i) = \frac{1}{N-1}\sum_j^N D_c(O_i, O_j)D_s(O_i, O_j)$ for each object O_i. D_c and D_s denote the color Euclidean distance and the normalized spatial Euclidean distance (the diagonal image length equals to 1) between objects, respectively.

The maximum distance between objects and fixations was set to 80 px which corresponds to about 2° of visual angle.

3.2.3 Experimental Results and Discussion

First, we explored saliency from color contrasts. We therefore compared the similarity between contrasts and the total number of fixations using the correlation coefficient (see Fig-

ure 3.6(a)). We found a strong correlation with both versions of color contrast which indicate that high-level factors have only a weak influence on attention.



(a) Correlations per image. (b) Correlations per participant.

Figure 3.6: Similarity between color contrasts and the number of fixations.

Furthermore, we looked into the individual effects of color contrasts on attention (see Figure 3.6(b)). Attention for low-level stimuli should be affected solely by bottom-up factors, but diverse scores support the individuality of color perception. For instance, attention of participant #8 is rarely directed towards contrast objects.

Finally, we investigated color prior by fixation ratios of each color (see Figure 3.7). Saliency from color contrasts seems to predominate over danger-links to colors in subjects' attention (see examples in Figure 3.8). Compared with a clear preference of red in the study of Gelasca et al. [67], we found only slightly higher fixations of red and yellow colors, therefore cannot confirm our hypothesis. This difference could be explained by fixations driven by unconscious attention. Etchebehere and Fedorovskaya [54] observed a discrepancy between the numbers of reported and fixated colors. Such a mismatch could also occur between reported saliency and selective attention.

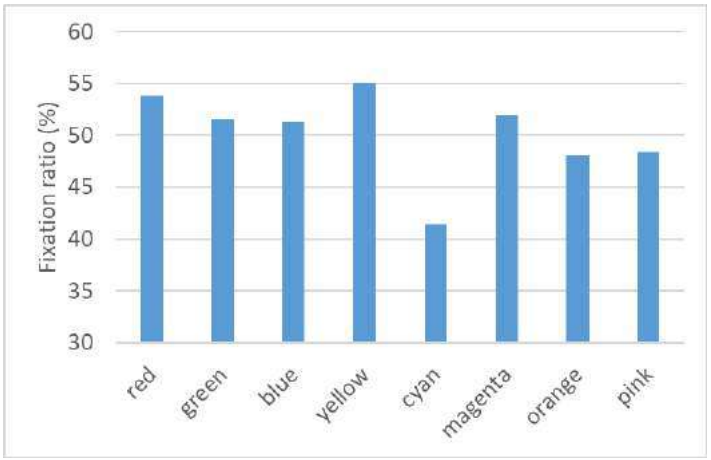
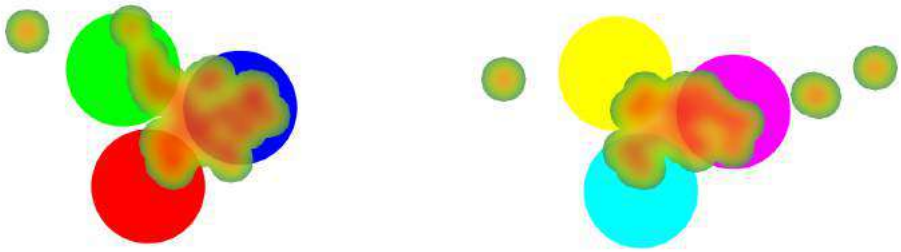


Figure 3.7: Fixation ratio of each color.

On the other hand, cyan objects were largely ignored with a remarkable gap in fixations. This could support the Wool et al.'s finding [212] that yellowish targets are detected faster than bluish ones.



(a) The most contrast object is blue. (b) The most contrast object is magenta.

Figure 3.8: Example fixation heatmaps where objects with the highest contrast are most salient. Warm colors (red and yellow, respectively) did not grab much attention in these scenes.

3.2.4 Summary

Our experiment confirms that saliency from color contrasts play an important role in attention. An unexpected observation was that the LAB color space could not equally estimate perceived color differences of all participants. Therefore, there are presumably other, memory-related factors, that color vision employs. However, we did not found a significant preference in fixating danger-related colors regardless of distractors. While there was only a negligible dominance for red and yellow, the experiment surprisingly showed significantly less fixations of cyan. Future experiments should therefore use more colors, more participants and other color spaces for a deeper investigation of the color perception individuality and psychological functioning.

3.3 Visual Attention to Shape

Computational models usually predict stimulus-driven human visual attention by simple feature saliency, such as intensity, color and orientation. Since object shapes and their contour segments influence attention too, this section investigates whether and to what extent global and local shape characteristics and their mutual differences affect visual saliency. To answer this question, we decided to employ diverse shape and contour descriptions and propose shape saliency models. To our best knowledge, an eye-tracking dataset focused solely on shape saliency has not been available so far. Therefore, we created such a dataset and made it publicly available. Recorded fixation data were used to identify frequently fixated shapes and evaluate proposed and existing shape saliency models. This work has been partially published in [235].

3.3.1 Motivation

Saliency models use shape information to predict visual attention only sporadically. However, human visual attention is also affected by object properties, such as object's size and shape [44]. Shape incorporation in a computational model could, therefore, improve saliency estimation. Available fixation datasets were usually recorded on natural scenes where various factors of attention are present. In order to model and evaluate shape saliency precisely,

we need to create a specialized image dataset that would suppress the effect of other attention aspects as much as possible. Since shape detection is preceded by contour detection [142], saliency model should take overall shape features as well as features of its contour fragments into account.

3.3.2 Experimental Study on Shape Saliency

We can assume that scenes with only object silhouettes is strongly guided by contours and shapes. We therefore conducted own eye-tracking experiment with various real and abstract silhouettes on a uniform background to analyze shape saliency.

To find out how shape affects attention, we used various techniques for contour and shape description and matching and evaluated them using recorded fixations.

Hypotheses

H1: *Salient object boundary has a different contour fragment from the neighboring object boundaries.* This assumption is included in most regional saliency models whose aim is to search for a global rarity. We expect that contour differences could lead in a significant pop-out effect because contour identification is performed in early stages of shape perception. We therefore developed location-based saliency models that compare contour fragments of objects. Each saliency value represents the average difference of the corresponding contour.

H2: *Salient object has a different global shape from the neighboring objects.* This hypothesis continues in H1 at object level. Hence, we proposed object-based saliency models that compare global shape descriptions of each object pair and assign a saliency value to each object proportionally to the average shape difference. In contrast to **H1**, we can use this approach in scenes containing at least three objects.

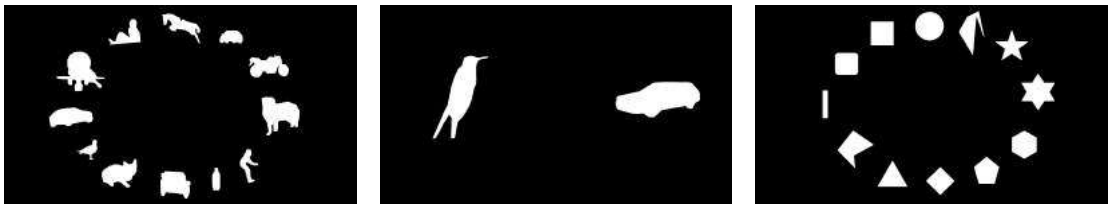
H3: *Some shape characteristics draw more attention ignoring across-shape similarities. Larger, asymmetrical and complex objects are more salient.* We reason that impending danger increases with object size. Attention whose original purpose is linked to protection is thereby shifted to larger figures that can potentially threaten us. Since high curvatures are attentive, we assume that complex objects could be very salient too. Furthermore, attention to complex shapes of abstract or rare objects could be affected in top-down manner due to their unexpected appearance which does not match to any familiar objects. Next, saliency of asymmetrical objects with many local curvature extrema could be partially related to **H1** due to many contour fragment differences that form the object and thereby attracts attention as the whole unit. Hence, we proposed object-based saliency models that consider objects with higher shape irregularity and size as more salient. These models are applicable to scenes with at least two objects.

In addition, we believe that all the above described effects are involved, to some extent, in visual attention.

Image Data

Since attention can be affected by many stimuli in natural images, we built a new dataset with own images focused on shape stimulus. Hence, we prepared 158 scenes with 208 object sil-

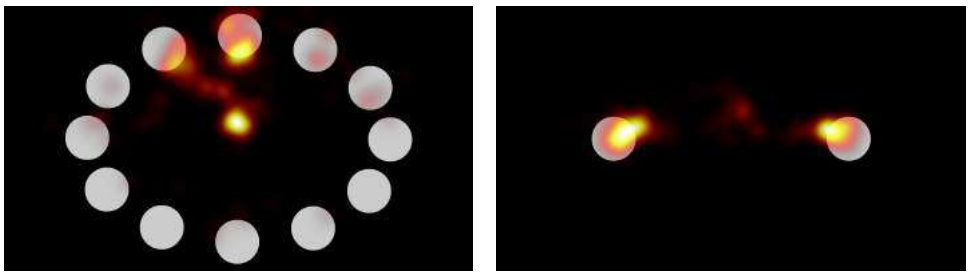
houettes, in total. We designed own abstract shapes and selected real-world objects from a standardized dataset of natural images for object class recognition provided by the Visual Object Classes Challenge 2012 [55, 56] which is considered as the representative object sample including persons, animals, vehicles and household articles. Each scene contains binary masks in a form of silhouettes on an uniform background, either 12 shapes organized in a circle (85 images) or 2 shapes on both image sides (73 images) (see examples in Figure 3.9).



(a) 12 natural objects. (b) 2 natural objects. (c) 12 geometric shapes.

Figure 3.9: Examples of object silhouettes displayed in our experiment.

In order to minimize the effect of reading patterns (from top to bottom, from left to right) [91] (see test images in Figure 3.10), we used the same objects multiple times at different positions, e.g. swap objects in two-object-scenes.



(a) 12 objects. (b) 2 objects.

Figure 3.10: Heatmaps from participants' first 4 fixations for test images with same object shapes. Heatmaps indicate that participants started to scan the scene from top and left, respectively.

Participants and Apparatus

We recorded gaze of 73 students participating a human-computer interaction course (65 males, 8 females). Their age ranged from 19 to 25 years. All subjects gave their informed consent to the study and received an explanation of the experiment. Their participation was compulsory to gain all credits for the course.

Fixations were recorded using Tobii X2-60 eye-trackers at 60 Hz. Images were displayed on 24.1-inch monitors with a resolution of 1920 $\times$ 1080 pixels at viewing distance of approximately 60 cm. Eye-tracking data were processed by Tobii I-VT fixation filter.

Experimental Design and Procedure

Participants were instructed to freely view scenes, each displayed for a fixed period of time (10 sec. for 12 objects, 5 sec. for 2 objects). Each image was preceded by a central fixation

cross. Participants were shown 43 up to 76 images.

Measures and Analysis

We took first 10 fixations of each participant to cover on the early stage of visual attention.

To define how shape saliency influences attention, we computed a relative number of fixations of each object, i.e. *fixation frequency* (fixations outside object regions were excluded). Then, we measured the *fixation-shape similarity* and the *fixation-saliency similarity* as the correlation coefficient (CC) between object fixation frequency and the corresponding value of shape description and shape saliency, respectively. To compute the normalized similarity to saliency maps, the maximum saliency values were normalized.

Furthermore, location-based saliency models (see Subsection 3.3.3) were compared with blurred fixation maps (Gaussian filter: size = 200, $\sigma = 32$) using the SIM score (see definition of the score in Section 2.10). The background regions of scenes were excluded in the SIM calculation.

Data of our experiment are available online[4].

3.3.3 Proposed Computational Models

To investigate shape saliency, we proposed three types of computational models that follow one of our hypotheses. The models employ shape descriptors and matchers to detect salient shapes (object-based models) or salient boundary contours (location-based models). Saliency is computed regardless of other objects or as the difference of a target object from non-targets.

Intra-Shape Saliency Models

The first type of models are based on the hypothesis **H3** so that saliency of each object is defined only by its global geometrical properties ignoring the spatial context. In other words, saliency is not modulated by visual properties of other objects.

We used simple shape features that globally describe a region by a single value. Since these features represent roughly geometrical object properties, they can discriminate only shapes with larger differences. Predicted saliency of each object is proportional to a value of simple global shape descriptors.

We expect a fixation preference for larger objects. Therefore, the following models assign higher saliency to objects with higher values of shape descriptors defined as *area size* (denoted **A**), *perimeter length* (denoted **P**) and *equivalent diameter*[5] (denoted **ED**) [222, 216].

We also expect that asymmetrical geometrical shapes attract visual attention. We proposed saliency models that predict higher saliency for objects with higher values of *eccentricity*[6] (denoted **EC**) and *aspect ratio*[7] (denoted **AR**) [222, 216].

[4] `http://vgg.fiit.stuba.sk/2018-07/shapeSal/`

[5] Equivalent diameter is the diameter of the circle whose area equals to the object area.

[6] Eccentricity represents the ratio of major and minor axis length.

[7] Aspect ratio is the ratio of length to width of bounding rectangle.

Furthermore, we consider irregular and highly curved shapes as salient. Hence, the following models assume that objects with higher *extent*[8] (denoted $\mathbf{EX^{-1}}$), *rectangularity*[9] (denoted $\mathbf{R^{-1}}$), *solidity*[10] (denoted $\mathbf{S^{-1}}$) and *circularity*[11] (denoted $\mathbf{C^{-1}}$) [222, 216] are less salient.

Inter-Shape Saliency Models

Saliency predicted by the second group of models is based on its unique global appearance in comparison to non-targets following the hypothesis **H2**. Object saliency is implemented as the average distance between the global shape descriptions.

First, we measured the difference of shape features used in intra-shape saliency models (these saliency models are denoted with suffix **_DIST**).

The next model is based on *Hausdorff distance* [222] of a target from non-targets (denoted **HD**). Objects are aligned by their centroids to obtain translation invariance. The distance between objects A and B is calculated as: $\max(h(A,B), h(B,A))$, where $h(A,B) = \max_{a \in A} min_{b \in B} \|a - b\|$, a and b represent boundary points.

The next computational model employs *shape context* algorithm [11] to describe and match shapes (denoted **SC**). For each boundary point, we find its vectors to all other boundary points. The point is then represented by a log-polar histogram of vectors' length and orientation, called the shape context. Shape matching is done by minimizing the total cost of matching the boundary points' shape contexts.

We also proposed a saliency model that represents a shape by a one-dimensional function, called shape signature, particularly *centroid distance*. The signature of boundary points distances to the centroid is normalized to being scale invariant. The closest point to the centroid is considered as a starting point of the signature. The model considers shapes with distinct signatures as salient. Therefore, it compares the signatures using one of 4 different metrics – correlation (denoted **CD_COR**), chi-square (denoted **CD_CHI**), intersection (denoted **CD_INT**) or Bhattacharyya distance (denoted **CD_BD**) to compute shape saliency.

Centroid distance is also used in a saliency model that analyzes the shape in spectral domain (denoted **FCD**). First, we apply discrete Fourier transform on the signature. The coefficients of the transform are also called *Fourier descriptors* [223]. To obtain scale invariant description, the coefficients' magnitude is divided by the DC component. Saliency between 2 shapes is computed as the Euclidean distance between their vectors of normalized Fourier descriptors.

Finally, we proposed a model based on *boundary moments*. The spatial moments are defined as $m_{pq} = \sum_x \sum_y f(x,y)$, where $p, q = \{0, 1, 2, ...\}$. To represent a shape contour, we used 7 geometric moment invariants [87] that are invariant under translation, scaling and rotation. The model predicts high saliency values to shapes whose moment invariants differ from other shapes. Corresponding moment invariants of two shapes h_A and h_B are compared as $|\frac{1}{m_A} - \frac{1}{m_B}|$ (denoted **HU1**), $|m_A - m_B|$ (denoted **HU2**) or $|\frac{m_A - m_B}{m_A}|$ (denoted **HU3**), where $m_i = sign(h_i) \cdot \log(h_i)$ [107].

[8] Extent represents the ratio of object area to its bounding rectangle.

[9] Rectangularity represents the ratio of object area to its rotated bounding rectangle.

[10] Solidity represents the ratio of object area to its convex hull.

[11] Circularity represents the ratio of object area to its perimeter square.

Contour Saliency Models

In contrast to object-level saliency of shape saliency models, contour models based on the hypothesis **H1** compute saliency at each contour point. We assume that attention directs human gaze to boundary points that differ from their surrounding contours.

Following the *center-surround* approach of Itti et al.'s model [96], we build a Gaussian pyramid from the centroid distance signature for our local shape model (denoted **CSCD**). It computes differences between center ($c = \{0, 1, 2, 3, 4\}$) and surround ($s = c + \delta$; $\delta = \{3, 4\}$) scales of the pyramid to detect salient contour points:

$$S(c, s) = |S(c) - S(s)| \tag{3.3}$$

Saliency value of each contour point is spread to a triangular area defined by the contour point, the subsequent contour point and the object centroid. To smooth discontinuities between these areas, a Gaussian filter is applied.

The next local model applies a Fourier transform on the centroid distance signature (denoted **SRCD**). It is based on a work of Hou and Zhang [86] that introduced the *spectral residual* approach to create a saliency map. Saliency of object boundary is then obtained by the inverse Fourier transform smoothed by a Gaussian filter. Computed saliency values are spread in the whole object using the same triangular method as in the CSCD model.

3.3.4 Experimental Results and Discussion

We used fixation data from our dataset to evaluate the prediction ability of our saliency models. In addition we evaluated intensity (denoted with suffix **_I**) and orientation saliency (denoted with suffix **_O**) of the center-surround model by Itti et al. [96] (denoted **ITTI**, implementation by Harel [80]) and the graph-based model by Harel at el. [80] (denoted **GBVS**) and the inter-shape saliency model by Chen et al. [37] (denoted **JI**). Figure 3.11 contains example saliency maps of all evaluated models.

Influence of Object Size, Asymmetry and Complexity

To test whether visual attention is directed to larger, asymmetrical and complex objects (**H3**), we measured fixation-shape similarities of intra-shape saliency descriptions for two- and twelve-object scenes, respectively. We therefore expected positive correlations for object size descriptions (*A*, *ED* and *P*) and negative correlations for shape characteristics describing object symmetry or complexity (*AR*, C^{-1}, *EC*, EX^{-1}, R^{-1} and S^{-1}).

Table 3.4: Fixation-shape similarity of simple shape properties.

Objects		**A**	**AR**	$\mathbf{C^{-1}}$	**EC**	**ED**	$\mathbf{EX^{-1}}$	**P**	$\mathbf{R^{-1}}$	$\mathbf{S^{-1}}$
2	CC	**0.39**	-0.17	-0.13	-0.24	**0.40**	-0.11	**0.31**	-0.05	-0.05
	p	< .001	.046	.113	.003	< .001	.197	< .001	.516	.563
12	CC	0.29	-0.20	**0.30**	-0.23	**0.30**	0.20	**0.54**	0.23	0.29
	p	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001

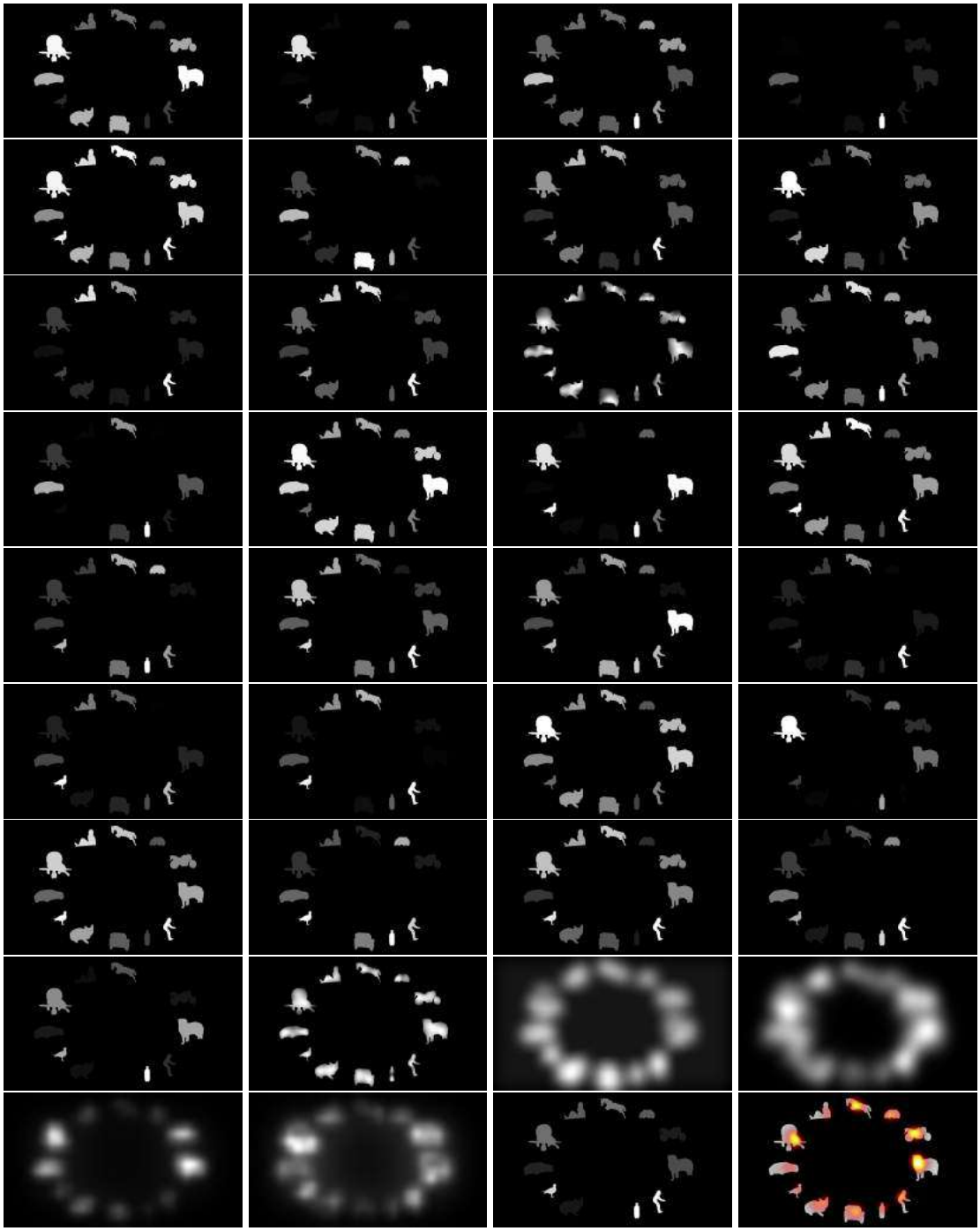
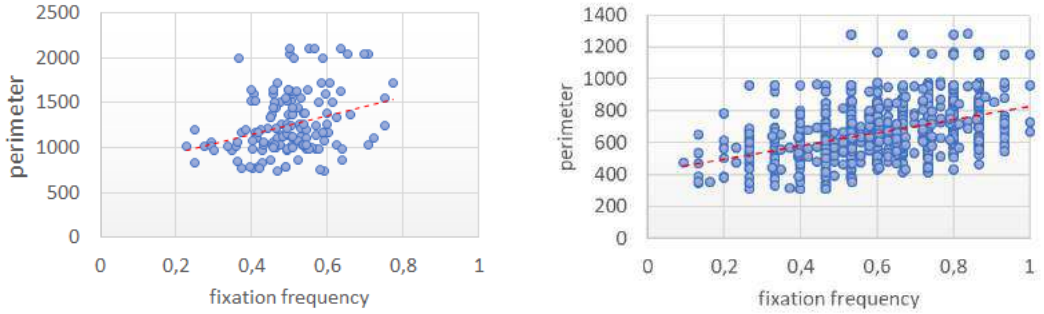


Figure 3.11: Saliency maps and a fixation heatmap of the scene shown in Figure 3.9 (left). From left to right, the 1st row – A, A_DIST, AR, AR_DIST; the 2nd row – C^{-1}, C_DIST, CD_BT, CD_COR; the 3rd row – CD_CHI, CD_INT, CSCD, EC; the 4th row – EC_DIST, ED, ED_DIST, EX^{-1}; the 5th row – EX_DIST, FCD, HD, HU1; the 6th row – HU2, HU3, P, P_DIST; the 7th row – R^{-1}, R_DIST, S^{-1}, S_DIST; the 8th row – SC, SRCD, ITTI_I [96], ITTI_O [96]; the 9th row – GBVS_I [80], GBVS_O [80], JI [37] and the fixation heatmap.

As listed in Table 3.4, we found a correlation between descriptions and fixations, but it is stronger only for objects defined by perimeter length for 12-objects-scenes. Though the results suggest that there could be a fixation preference for larger objects (see *A*, *ED* and *P* in Figure 3.12), we obtained either negative or ambiguous correlations across both scene types

for other shape descriptions regarding to object asymmetry and complexity. While a positive correlation for of the $\mathbf{C^{-1}}$ supports **H3** so that salient shapes deviate from circles, a negative correlation of the *EC* model could indicate the opposite to what we expected – saliency of symmetrical shapes.



(a) 2 objects. (b) 12 objects.

Figure 3.12: Relationship between perimeter lengths of objects and their fixation frequency. The red dashed lines are lines of best fit.

Since we did not observe a clear trend for fixating asymmetrical and complex objects, we reject the hypothesis H3 as a whole and restrict this effect only to the object size.

Influence of Global Shape Rarity

Next, we investigated whether global rarity saliency could be also associated to object-level shape (**H2**).

We therefore analyzed the performance of inter-shape saliency models and the *JI* [37] model which define saliency using the dissimilarity to other shapes in an image (this saliency could be predicted only in 12-objects-scenes).

We analyzed the normalized fixation-saliency similarity. Among saliency models that employ simple descriptor differences, we only found a weak correlation between shape saliency and participants' fixations for the model based on the perimeter feature (*P_DIST*), but it is weak ($CC : 0.21; p < .001$). The rest of inter-shape saliency models did not correlate with fixations, but the *CD_COR* model whose similarity is low too ($CC : 0.21; p < .001$).

A possible explanation could be that this process requires an analysis of all contour and shape characteristics so that identification of an uniquely shaped object takes much longer time than intra-shape saliency and thereby contrast saliency does not result in immediate pop-out. *We concluded that global shape rarity does not strongly correlate with human fixations. Therefore, we cannot confirm the hypothesis H2.*

Influence of Contour Segment Differences

To test whether attention is affected locally by differences of contour segments (**H1**), we evaluated our contour saliency models as well as *Itti* [96] and *GBVS* [80] models using the SIM method. Even though *Itti* and *GBVS* do not explicitly compute contour differences, center-surround intensity and orientation contrasts could contribute to shape perception too.

As visualized in Figure 3.13, all evaluated saliency models achieved a strong similarity to fixation maps (see example saliency maps in Figure 3.14). The repeated measures ANOVA with Bonferroni-adjusted posthoc comparisons showed that saliency predicted by the *SRCD* model is significantly more similar to participants' fixation maps than the rest of location-based saliency maps ($F(3.774, 592.40) = 27.52; p < .001; \eta^2 = .149$; Greenhouse-Geisser correction)[12].

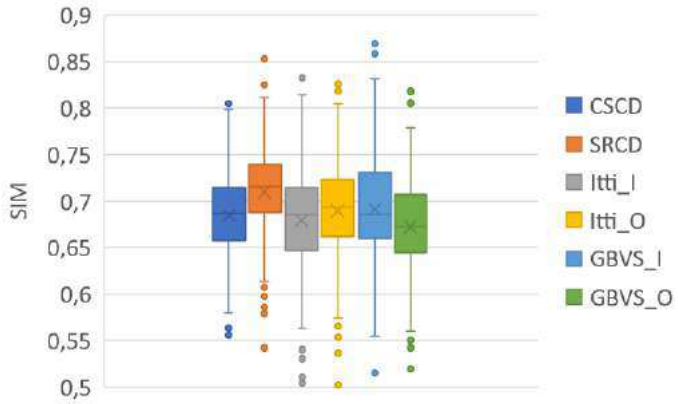
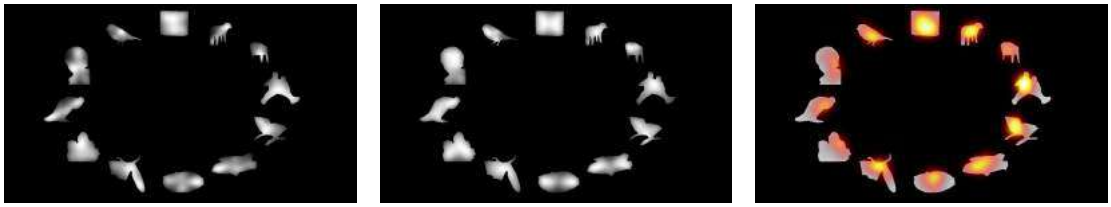


Figure 3.13: SIM scores of location-based saliency models.



(a) CSCD. (b) SRCD. (c) Heatmap.

Figure 3.14: Saliency predicted by proposed contour models and a fixation heatmap.

Even though intensity and orientation differences can only roughly describe object boundaries, attention is significantly affected by these contrasts. The method based on differences of the centroid distance in the spectral domain (SRCD) achieves considerable highest prediction accuracy for shape saliency. Unique boundary object parts play an important role in visual attention, thereby confirming hypothesis H1.

[12]**Analysis of variance** (ANOVA) is a statistical method which should be used to determine differences between means of several groups instead of multiple **t-tests**. ANOVA assumes independence of observations, normal data distribution within each group and homogeneity variances across the groups. If the assumptions are violated, the non-parametric equivalent, **Kruskal-Wallis H** test should be utilized instead (non-parametric **Mann-Whitney U** test could be used to compare two groups). A one-way ANOVA considers only a single independent variable to compare the groups. In contrast, the **repeated measures ANOVA** (RMANOVA) is used for related groups instead of independent ones. In other words, there is the same group of participants tested multiple times or under different conditions. One of the RMANOVA assumptions is sphericity. If it is violated, the *Greenhouse-Geisser* and the *Huynh-Feldt* corrections can correct the statistic. The non-parametric version of RMANOVA called **Friedman test** should be considered if parametric assumptions are violated (non-parametric **Wilcoxon signed-rank test** test could be used to compare two related groups) [45, 76, 61].

High-Level Factors of Shape Saliency

Saliency models presented in this section aimed to describe and distinguish only low-level shape properties. However, attention is guided by high-level factors too, including memory representations of objects, preferred shape types such as human faces and categorical relationships between objects.

Indeed, Figure 3.15 illustrates objects with fixations concentrated at animal and human heads. In addition, human-like figures seems to be often more salient than silhouettes of non-living objects (see Figure 3.16). This finding may refer to the original aim of visual attention, to warn of impending danger which could be associated with living beings.

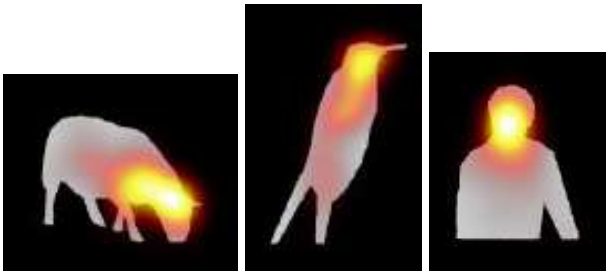


Figure 3.15: Participants frequently fixated animal and human heads.

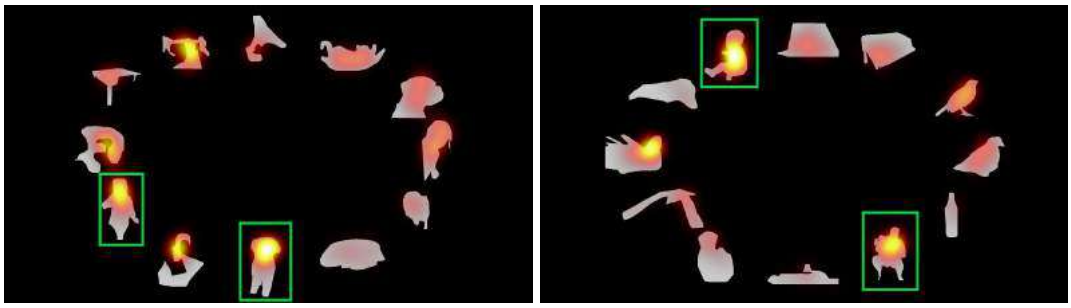


Figure 3.16: Participants' gaze was often directed towards human-like figures (marked with green rectangles).

3.3.5 Summary

Our results indicate that shape information could increase the accuracy of standard saliency models. While we showed that users tend to focus on larger objects, we did not find the increased attention in irregular and complex objects for both, 2- and 12-objects-scenes. Furthermore, we showed that global shape contrasts have only a negligible effect on visual attention. On the other hand, local boundary contour contrasts significantly grab users' attention. The proposed spectral residual approach applied on a centroid distance signature models contour saliency with the highest precision.

In addition to bottom-up factors, attention in our experiment seems to be modulated in top-down manner, by the semantics of object silhouettes. We observed various scenes where humans and faces tend to be fixated more often than other object categories. In order to cover these high-level factors of shape saliency, we could use a neural network trained on human fixations.

In the future, we should extend the experiment with varying number of displayed objects to find out whether the influence of global shape rarity increases in scenes with a lower number of objects. Also, gaze behavior in real scenes needs to be explored to compare saliency of shape to other features.

3.4 Visual Attention to Egocentric Depth

Computational models usually do not employ depth information to estimate saliency. However, depth is an important aspect of visual attention in egocentric vision. This section explores how the relative distance of objects affects attention using own eye-tracking experiments. In contrast to previous studies that investigated depth on 2D and 3D images, our experiments whose fixation data are publicly available took place in a natural environment. The main study outcomes are derived relationships between stereoscopic depth and depth contrasts of real objects and saliency which could be used in egocentric attention modelling.

3.4.1 Motivation

As mentioned in Section 3.1, analysis of egocentric vision has gained increased interest with the increasing use of mass-marketed miniaturized wearable cameras. With this a person is taking visual measurements about the world in a sequence of fixations which contain relevant information about the most salient parts of the environment and the goals of the actor. Depth perception in a natural environment does not arise only from pictorial cues that may be depicted in images, but also by motion-produced cues and binocular cues [75]. A majority of prior experiments showed higher saliency for objects closer to a viewer using fixations on 2D and 3D images [98, 125, 202]. However, a research on 3D depth saliency in a natural environment with real objects that could differ from stereoscopic images has largely been ignored. If it shows a depth bias, we could use it to increase the prediction ability of computational models for videos captured from the wearer's camera.

3.4.2 Experimental Study on Depth Saliency

The studies exploring the importance of depth channel in attention focused on 2D and 3D images. Most of them concluded that fixations are biased towards areas close to a an observer [98, 125, 202]. They also observed the non-linear effect of depth on visual attention. Since previous works showed advantages of binocular vision in depth discrimination [5, 164], we performed an eye-tracking experiment to find out whether depth saliency has the same effects in a natural 3D environment as in image viewing conditions.

Subjects participating in our study were asked to freely view identical objects placed in an experimental room that varied in depth channel. Recording their fixations, we investigated the influence of object distance and depth contrasts on the fixation order and the fixation distribution.

Hypotheses

H1: *Objects relatively closer to an observer are more salient.* We assume that the depth bias observed for images is consistent with fixations in a real environment. We therefore expect that the closest objects are fixated more frequently (**H1.1**) and earlier than distant ones (**H1.2**).

H2: *The distance of salient objects differs from the neighboring objects.* This assumption is based on the pop-out effect that arises from distinct regions. This global rarity saliency is implemented in most regional saliency models, e.g. for intensity, color and motion channel. We therefore expect that objects distant from other ones are highly fixated (**H2.1**) and rapidly detected (**H2.2**).

Stimuli

In our study, we employed 8 identical white balls of 13 cm diameter. The balls were attached to wires that hang on a ceiling at different depth positions in the room, as visualized in Figure 3.17. To suppress the center-bias, objects were arranged in an octagonal layout which centers at a viewer's visual field (instead of objects arranged in a single line [233]). Objects were located at 6 different depth planes, starting from the distance of approximately 2.5 m up to 4 m, by a step of 30 cm.

We designed 5 scene types to study influences of depth and depth contrasts on visual attention. To minimize the effect of reading patterns (from top to bottom, from left to right) [91], we prepared three variants of each scene type. Object depth levels (distances) are listed in Table 3.5. The number of different depth levels ranged from 2 up to 6 and the steps (differences) between adjacent object depth levels varied from 0 up to 4.

Table 3.5: Object layouts. Each layout is presented by object depth levels, from **A** to **H** (object labels and depth levels are defined in Figure 3.17(c) and 3.17(d), respectively). For example, the layout **V/b** corresponds to Figure 3.17(b).

Scene type	**Depth levels**	**Depth level steps between adjacent objects**	**Variant**		
			a	*b*	*c*
I	2	the only one step of 1	12111111	11121111	11111112
II	3	alternating zero steps and steps of 2	33553311	53311335	31133443
III	4–5	only steps of 1	34565654	32123454	56543234
IV	4–6	steps of 1, but one step of 3	45632123	52321234	54565434
V	5–6	steps vary from 0 up to 4	62543134	32456321	34662124

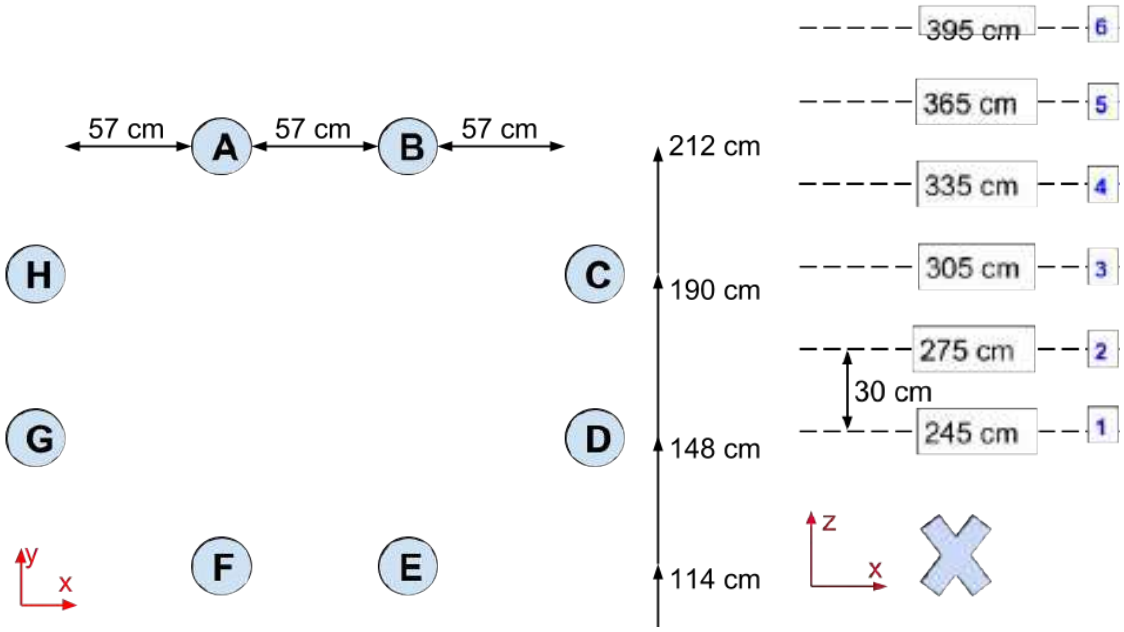
Experimental Design and Procedure

Participants were shown one of three scene variants for each object layout (see Table 3.5) with no specific task instruction.

Before they saw a prepared scene, they were instructed to stand on a fixed position horizontally aligned with the center of the octagonal layout while they kept their eyes closed.

(a) Experimental room. A participant freely looked at identical balls organized in an octagonal layout.

(b) Observer's view of a scene. Participants were initially instructed to stand on a fixed position and direct their gaze at the layout centre.



(c) Frontal view of a scene. Balls (labelled by letters **A** to **H**) are placed at 4 different horizontal and vertical positions.

(d) Top view of a scene. There are 6 distance levels from a viewer ($\times$), labelled from **1** (closest) to **6** (farthest).

Figure 3.17: Experimental setup.

In addition, they were navigated to move their eyes approximately at the layout center, as shown in Figure 3.17(b) .

The whole experiment took about 10 minutes per participant, on average.

Measures and Analysis

We recorded users' fixations on balls and further analyzed only first 10 sec of viewing (fixations outside object regions were excluded). We measured fixation ratio of each object and order of object visits.

Relative depth of objects refers to normalized depth object value where the minimum (0) and maximum (1) values correspond to the closest and farthest object distances in a scene, respectively. In addition, we represent each object by *relative depth difference* defined as the average difference with relative depths of all other objects in a scene.

Eye-tracking data of our experiment are publicly available[13].

Apparatus

We recorded participants' gaze using SMI ETG 2 eye-tracking glasses at 60 Hz. Gaze data were processed by Tobii I-VT fixation filter. Participants could perform any head movements while they stood still in a fixed position. The experiment was conducted in artificial lighting conditions.

Participants

We collected gaze data of 28 students participating a human-computer interaction course aged from 19 to 25 years; 25 males and three females. All subjects gave their informed consent to the study and received an explanation of the experiment. Subjects who normally wore glasses or contact lenses for distant viewing were asked to wear them during the experiment. Their participation was compulsory to gain all credits for the course.

3.4.3 Experimental Results

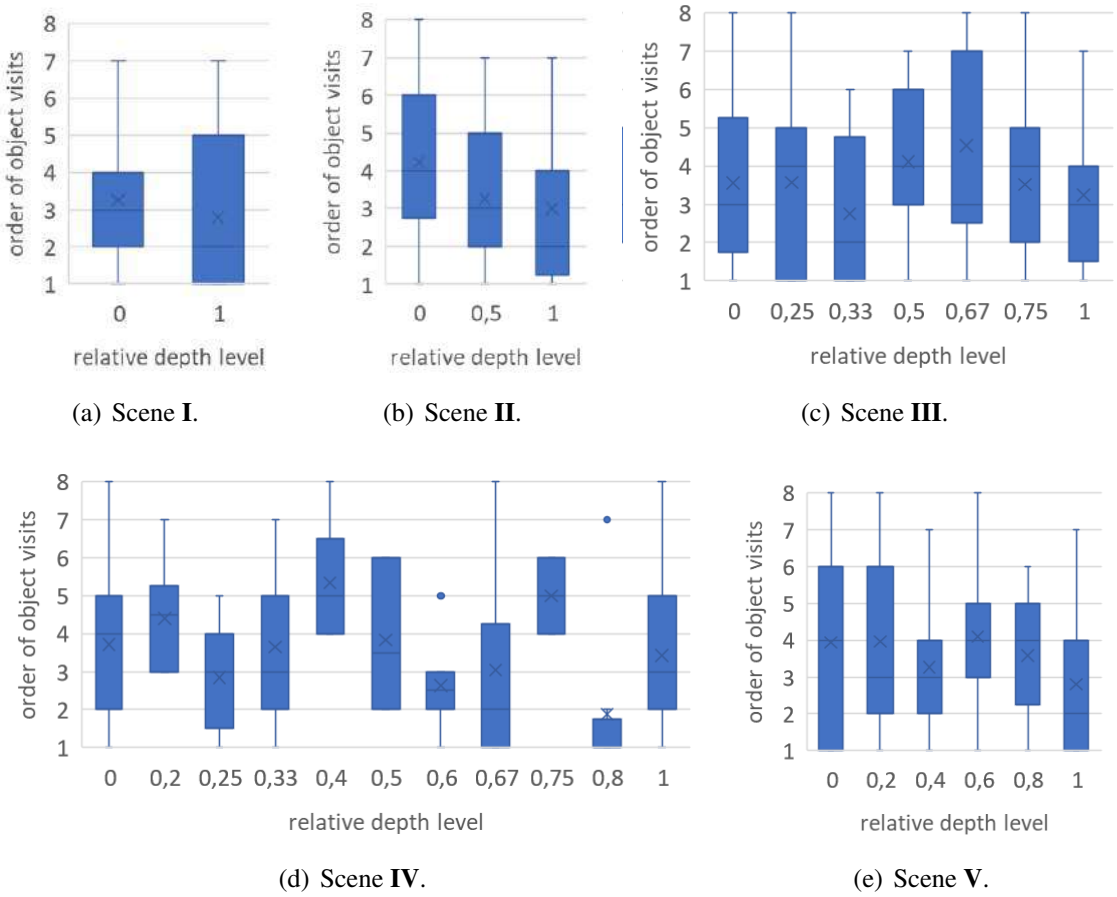
We evaluated participants' first fixations on 8 objects in 15 scenes (variant *a*: 9; variant *b*: 7; variant *c*: 12 participants). On average, participants fixated each object for 473 msec in total. Users visited at least once 66.4% of all presented objects in the room, resulting in 5 visited objects per scene, on average.

Order of Object Visits

First, we investigated the impacts of relative depth and depth contrasts on order of visited objects (**H1.2**). A Pearson correlation did not find a linear relationship between the object order and relative depth values. Furthermore, we checked the statistical significance in visits for each scene type separately. An ANOVA revealed statistically significant differences only in Scene **II** ($F(2, 133) = 3.722; p = .027; \eta^2 = .053$; Figure 3.18(b)) and **IV** ($F(10, 129) = 2.081; p = .030; \eta^2 = .139$; Figure 3.18(d))[12]. Surprisingly, the closest objects are fixated later than the farthest ones in Scene **II** (Bonferroni-adjusted posthoc comparisons). *Closer objects do not attract egocentric attention earlier. Thereby we have to reject the hypothesis H1.2*

Next, we analyzed whether the order of visited objects could be related to depth contrasts (**H2.2**). However, the only contrasted (farther) object in Scene **I** is not detected significantly sooner. ANOVA tests did not find any scenes with significant differences. *This means that depth contrasts do not affect the fixation order, thereby rejecting hypothesis H2.2.*

[13] `https://vgg.fiit.stuba.sk/2019-02/depthSal/`



(a) Scene **I**. (b) Scene **II**. (c) Scene **III**.

(d) Scene **IV**. (e) Scene **V**.

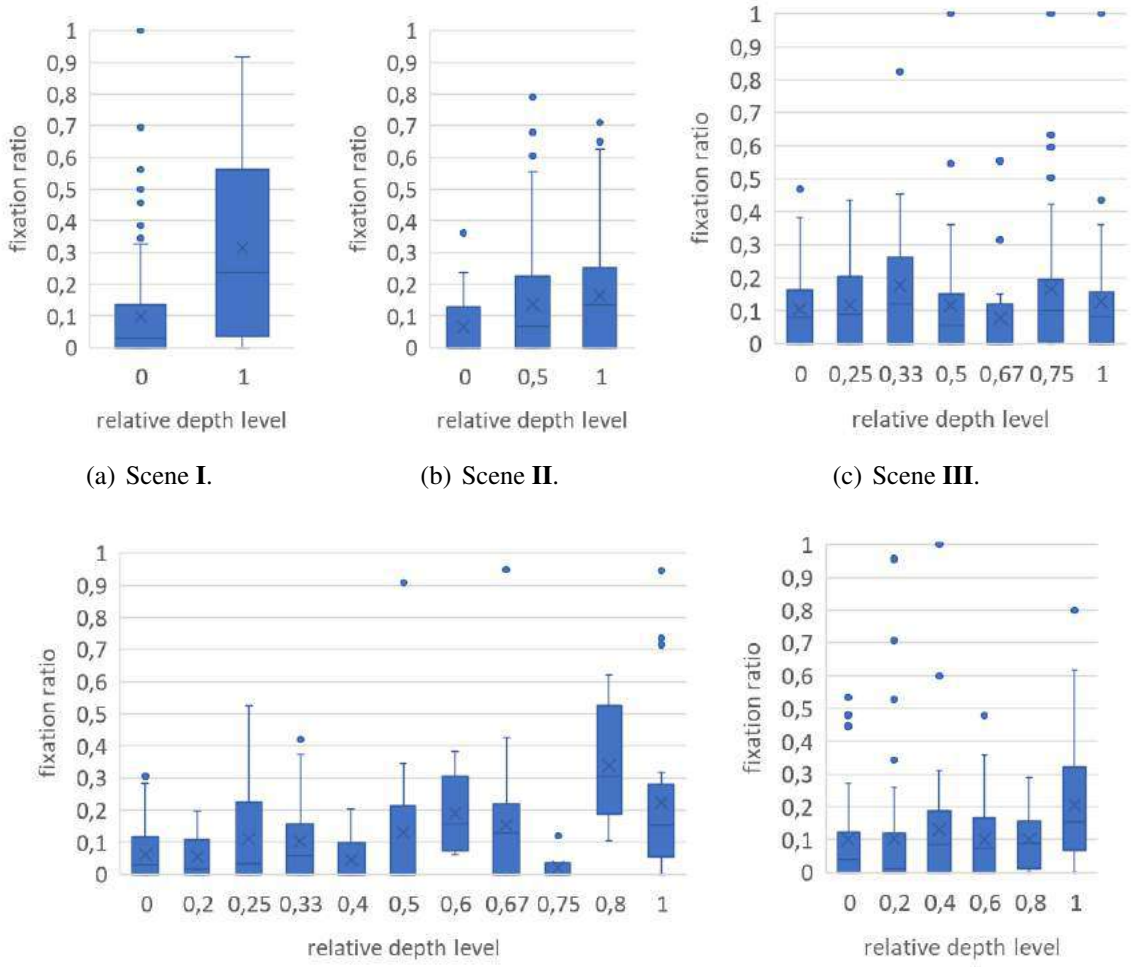
Figure 3.18: Order of object visits for each scene type grouped by object relative depth values (e.g. 1 = the 1st visited object).

Fixation Ratio of Objects

To test whether fixations are focused on closer objects (**H1.1**), we measured participants' fixation ratio for each depth level. We found a positive correlation between relative object depths and fixation ratio, but this correlation is weak ($r = 0.199; p < .001$). Visualizations of fixation ratios for each scene in Figure 3.19 indicate that the closest object to a viewer is not highly fixated, but there is a bias towards the farthest regions (except for Scene **III** which is the only scene without any significant differences using a Kruskal-Wallis H test). Another unexpected finding is a local maximum for fixations in the 3rd depth level (see Scene **III**, **IV** and **V**).

Grouping all scenes together, a Kruskal-Wallis H test indeed confirmed statistical significant differences in fixation ratios ($\chi^2(10) = 54.96; p < .001$; Figure 3.21(a)). Dunn's posthoc test showed that the farthest depth level (1) is fixated most significantly, except for an adjacent depth level of 0.8. *In other words, users focused on farthest objects, the opposite to what we assumed. Therefore we have to reject hypothesis H1.2.*

Furthermore, we analyzed whether depth contrasts affect fixations too (**H2.1**). We found only a very weak positive correlation ($r = 0.145; p < .001$), but there is an evidence for strong biases towards the most contrast objects only for Scene **I** and **V**, as shown in Figure 3.20. On the other hand, Kruskal-Wallis H tests did not find any significant differences for Scene



(a) Scene **I**. (b) Scene **II**. (c) Scene **III**.

(d) Scene **IV**. (e) Scene **V**.

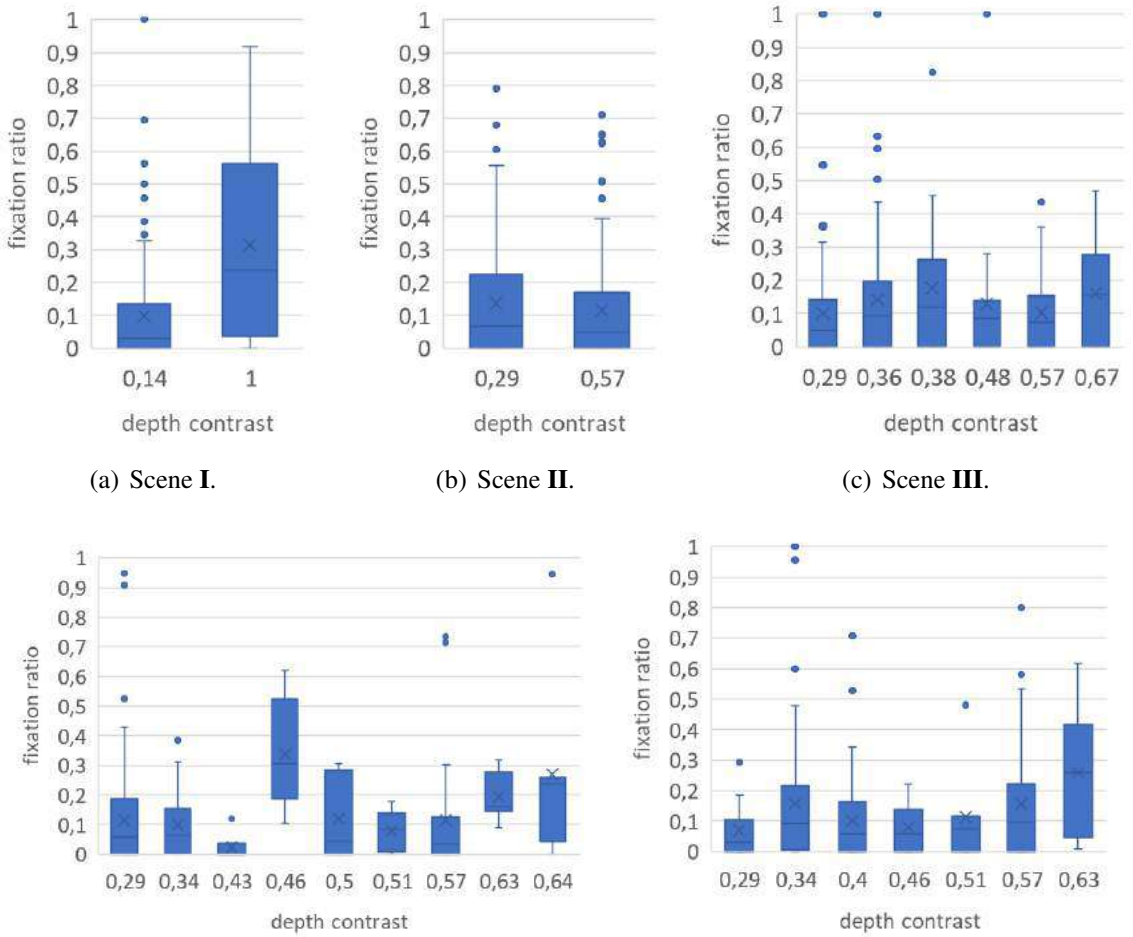
Figure 3.19: Distribution of participants' fixations for each scene type grouped by object relative depth levels.

II and **III**.

Considering all scenes, from a Kruskal-Wallis H test followed by Dunn's post hoc tests it could be concluded that the maximum depth contrast of 1 is fixated significantly higher than contrasts below a value of 0.63 ($\chi^2(15) = 50.37; p < .001$; Figure 3.21(b)). *In other words, high-contrast objects in depth channel grab users' attention even though this effect is not linear, thereby confirming hypothesis H2.1.*

3.4.4 Discussion

Representing objects by their depth level, we found that closer objects are not fixated first (**H1.2**) nor more frequently (**H1.1**). Furthermore, our experiment revealed that users' attention is more strongly directed towards distant objects in a real environment. Even though an experiment performed by Ramasamy et al. [173] showed that deeper parts of scenes could be more salient than the closer ones, our observation contrasts with a majority of prior studies exploring depth in image data. This inconsistency could indicate a different gaze behavior between real and stereoscopic image depths. Conventional saliency models designed for



(a) Scene **I**. (b) Scene **II**. (c) Scene **III**.

(d) Scene **IV**. (e) Scene **V**.

Figure 3.20: Distribution of participants' fixations for each scene type grouped by object depth contrasts.

pictorial data are thereby not suitable for egocentric video sequences.

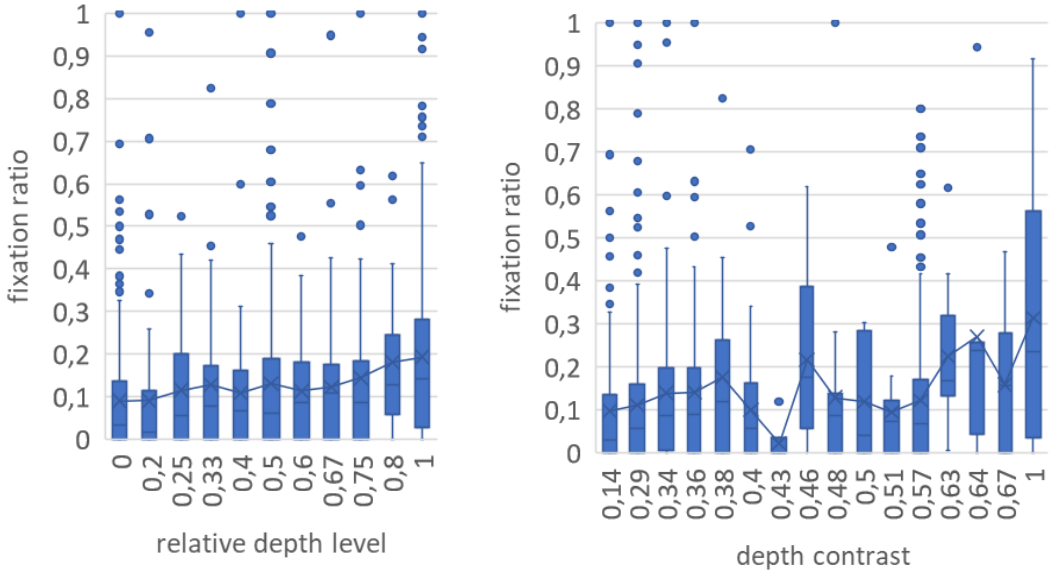
Our analysis of depth contrasts revealed that contrast objects influence fixation distribution (**H2.2**), but not their order (**H2.1**). The effect on fixations was expected and already implemented in region-based contrast models. However, in contrast to these computational models, the observed effect of depth contrast is not linear.

To predict users' attention in real environments, we measured fixation distributions on objects for each scene layout in our experiment and from the average values derived non-linear relationships between object distance to a viewer and other objects in a scene and saliency, as visualized in Figure 3.4.4 and 3.4.4, respectively. Fitting polynomial functions to the data, we obtained the following equations for both effects:

$$D(i) = 0.2394i^6 - 4.2447i^5 + 9.3252i^4 - 7.2134i^3 + 2.0952i^2 - 0.1068i + 0.0943, \quad (3.4)$$

$$DC(j) = 0.2891j^3 - 0.1443j^2 + 0.0576j + 0.1027, \quad (3.5)$$

where D and DC represent depth and depth contrast saliency, i is a relative (normalized) object depth value and j equals to the average distance to relative depth levels of other objects in a scene. The both equations could be used as an additional weighting factor of



(a) Relative depth.

(b) Depth contrast.

Figure 3.21: Distribution of participants' fixations on depths and depth contrasts showed a focus on farther and contrast objects.

traditional saliency maps.

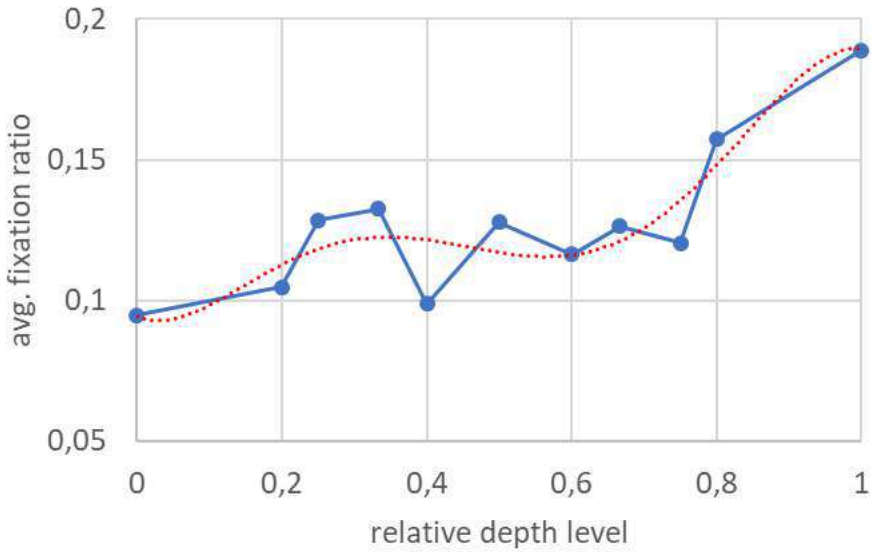


Figure 3.22: Predicted saliency of normalized object distance to a viewer. The red dashed lines represent a best polynomial fit.

There are some limitations of our experiment. Even though we tried to minimize the effect of other objects on visual attention in the experimental room, objects in peripheral vision, such as tables, lamps and windows could affect users' attention.

Next, a maximum object distance in our study was set to 4 m. Despite of a depth bias to farther objects we believe there is a threshold distance when object saliency starts to decrease since their size is relatively too small to grab attention. Therefore, the derived depth saliency

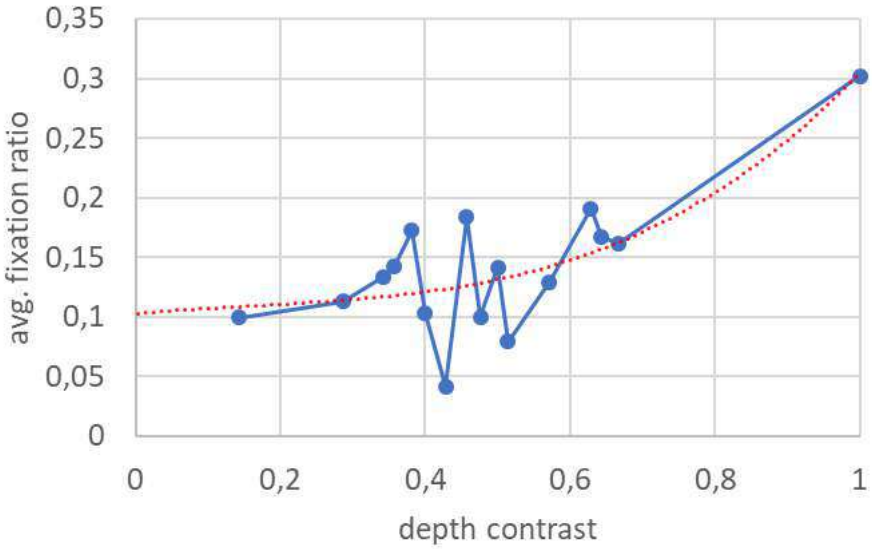


Figure 3.23: Predicted saliency of depth contrast with other objects. The red dashed lines represent a best polynomial fit.

is limited only to smaller object distances. In the future, it will be therefore important to explore saliency at larger distances.

Another aspect of attention which we did not analyze is a perceptual distance between objects. For instance, Scene **II/b** and **II/c** consist of 3 object pairs at different depth levels while the horizontal distance of each pair remains constant. However, objects at the closest depth level are seen as being horizontally the most distant pair. Hence, future studies could examine the effect of perceptual depth contrast instead of real depth contrast.

3.4.5 Summary

Our study explored the impact of depth saliency on egocentric vision in a real environment in contrast to prior depth experiments analyzing pictorial data. The results revealed a non-linear impact of object distance to a viewer and other objects on fixations in an environment. While attention on 2D and 3D images is directed towards closer objects, we found a bias towards more distant objects in our experiment. Therefore, we conclude that human gaze behavior in real environments differs from stereoscopic image displays and standard 3D saliency models incorporating depth channel are not suitable for egocentric video sequences. In addition, our experiment showed that depth contrast objects influences users' attention too. Based on our data we suggested depth saliency extensions for computational models in real environments.

3.5 Static Feature-Based Egocentric Visual Attention

Our experiments described in previous sections (Section 3.2, 3.3 and 3.4) explored the effects of static features on attention separately. We have already pointed out that binocular cues enhance viewer's perception (see Section 3.4). To find out how static stimuli compete for our attention in everyday actions, we conducted an eye-tracking experiment in a real environment

recorded from the first-scene perspective. To cover all aspects of human egocentric vision, we captured scene depth too.

Our analysis is focused on the relationships between saliency and static stimuli such as intensity, color, orientation, depth, object shape and viewer's central visual field. We examined how each feature influences human gaze using feature saliency maps predicted by various computational models. We proposed novel approaches to predict feature saliency and compared their performances with existing saliency models. In addition, we discussed the individuality of egocentric visual attention. Our experiment has been partially published in [233].

3.5.1 Egocentric Experiment

In the last decades, visual attention research was concerned with 2D viewing condition. Recent studies, partially due to the advent of wearable gaze recorders, tend to focus on an egocentric perspective of attention. Yamada et al. [215] analyzed egocentric views of participants sitting on chairs looked around a room where another person was randomly walking. Results showed that static attentive stimuli of the first-person attention are estimated by classic saliency models with high accuracy, whereas dynamic stimuli cannot be predicted efficiently due to missing egomotion compensation. Therefore, Yamada et al. [214] enhanced saliency prediction by egomotion estimation. Visual attention prediction has been also utilized to recognize activities from the first-person perspective [151, 59]. Li et al. [132] proposed a model to predict egocentric gaze using viewer's hand locations, head and hand motion estimation. Zhang et al. [230] introduced a 3D convolution neural network architecture for egocentric gaze estimations. Li et al. [134] used supervised saliency maps based on sparse coding to predict human gaze from the first-person perspective. Tavakoli et al. [193] examined task-specific egocentric datasets. They found a poor performance of conventional saliency models that are significantly outperformed by deep neural networks. Their analysis also showed a strong center bias and the advantage of manipulation points identification instead of hand detection.

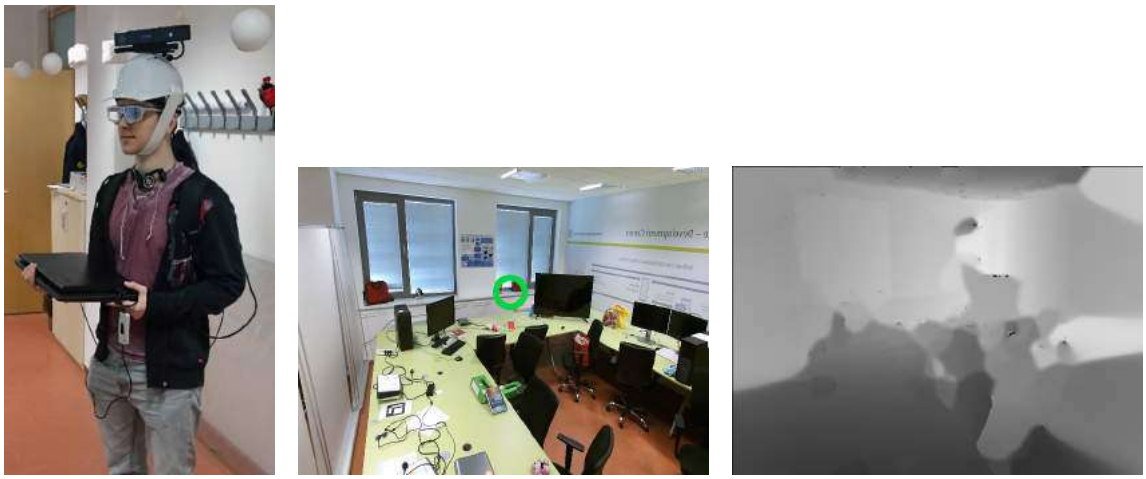
In contrast to previous research, we explored all fundamental low- and mid-level static factors that influence task-free binocular attention in a real environment including the effects of depth. We examined the ability of standard saliency models [96, 80] to predict feature saliency maps. In addition, we proposed novel methods to predict shape and depth saliency using our research findings that we report in Section 3.3 and 3.4, respectively.

Experimental Design, Apparatus and Participants

We conducted an eye-tracking experiment with 6 students who were asked to freely walk and explore a laboratory room for 15 up to 30 sec. All subjects voluntary participated in the experiment. They gave the informed consent to the study and received an explanation of the experiment.

The viewer's perspective and gaze were captured by SMI ETG 2 eye-tracking glasses at 60 Hz. The object distance was estimated by Kinect V2 device mounted on a participant's head as shown in Figure 3.24. Since our aim is to focus on the viewer's attention affected only by static stimuli (except for the effects of egomotion), the room did not contain any moving objects.

Since an eye-tracking egocentric dataset accompanied with depth information has not been available so far, we made recorded RGB-D video sequences publicly available[14]. Each video frame includes RGB-frame, depth frame and a participant's fixation[15].



(a) Participant's view was recorded by eye-tracking glasses and Kinect.

(b) Example video frame (fixation labelled by a green circle).

(c) Corresponding depth image.

Figure 3.24: Experimental setup.

Goals and Analysis

The primary goal of this experiment is to measure how the following static factors affect egocentric visual attention:

1. low-level features such as intensity, color, orientation and depth,
2. mid-level features such as object shape and contours,
3. viewer's central visual field (center bias).

We investigated whether the effects are reduced over time and the individuality of egocentric gaze behavior. In addition, we measured saccade lengths.

Feature saliency is represented by conventional [96, 80] and novel computational models that decompose overall attention to separate feature saliency maps. We analyzed which model is best suited for each feature of egocentric vision using NSS scores (see definition in Section 2.10)[16].

Feature Saliency Maps

Each feature in our study is represented separately by the following computational saliency models:

[14] `http://vgg.fiit.stuba.sk/2016-06/egocentric-rgb-d-eye-tracker-dataset/`

[15] Since both devices captured a scene at different frame rate, we had to synchronize their outputs. In order to register the images, we computed a homography using RANSAC from the corresponding SIFT [143] keypoints. Finally, we smoothed Kinect depth maps.

[16] Since each fixation is represented by small circular area, we used the maximum saliency value in the fixation region.

1. **Intensity:**
   - *center-surround* model by Itti et al. [96] (denoted **Itti_I**; Section 2.3),
   - *graph-based* model by Harel et al. [80] (denoted **GBVS_I**; Section 2.3),
   - our *superpixel-based center-surround* model [234] (denoted **SPX_I**; Section 3.1) which correlates superpixel histograms of intensity at finer and coarser Gaussian pyramid layers (the highest saliency is estimated for uncorrelated superpixels).
2. **Color:**
   - *center-surround* model by Itti et al. [96] (denoted **Itti_C**; Section 2.3),
   - *graph-based* model by Harel et al. [80] (denoted **GBVS_C**; Section 2.3),
   - our *superpixel-based center-surround* model [234] (denoted **SPX_C**; Section 3.1) which measures color distances between superpixels at finer and coarser Gaussian pyramid layers.
3. **Orientation:**
   - *center-surround* model by Itti et al. [96] (denoted **Itti_O**; Section 2.3),
   - *graph-based* model by Harel et al. [80] (denoted **GBVS_O**; Section 2.3),
   - our *superpixel-based center-surround* model [234] (denoted **SPX_O**; Section 3.1) which correlates superpixel histograms of orientation at finer and coarser Gaussian pyramid layers (the highest saliency is estimated for uncorrelated superpixels).
4. **Depth:**
   - *simple depth* model which linearly increases saliency with shorter object distance to a viewer so that the closest objects are most salient (denoted **D_linear**),
   - *experimental depth* model based on Equation 3.4 derived from our depth experiment (denoted **D_nonlinear**; Section 3.4),
   - *simple depth contrast* model which defines saliency linearly as global contrast of superpixels [1][17] (denoted **DC_linear**),
   - *experimental depth contrast* model which weights global contrast of superpixels[17] by Equation 3.5 derived from our depth experiment (denoted **DC_nonlinear**; Section 3.4),
   - our *superpixel-based center-surround* model [234] (denoted **SPX_D**) which correlates superpixel histograms of depth at finer and coarser Gaussian pyramid layers (the highest saliency is estimated for uncorrelated superpixels)[18].
5. **Shape:**[19]

[17] We defined global contrast as $S(r_i) = \sum_{i \neq j} D_d(r_i, r_j) exp(-D_s(r_i, r_j))$, where r_i denotes the i-th region, D_d is the distance between average region depths and D_s is the normalized Euclidean distance between region centroids.

[18] This is exactly the same approach as superpixel-based intensity saliency **SPX_I** [234] (see Section 3.1).

[19] Each shape saliency model employs superpixels [1] followed by DBSCAN clustering algorithm [53] to segment images [119]. DBSCAN is a density-based clustering based on the distance measure and the minimum number of points which works with a concept of noise. In our case, it utilizes the mean color distance between superpixels.

- *perimeter intra-shape* model that assigns higher saliency to larger regions defined by perimeter length (denoted **S_perimeter_intra**; Section 3.3),
- *perimeter inter-shape* model that measures global contrast of region perimeters[17] (denoted **S_perimeter_inter**; Section 3.3),
- *equivalent diameter intra-shape* model that assigns higher saliency to larger regions defined by equivalent diameter (denoted **S_eqdiameter_intra**; Section 3.3),
- *equivalent diameter inter-shape* model that measures global contrast of regions represented by equivalent diameter[17] (denoted **S_eqdiameter_inter**; Section 3.3),
- contour *CSCD* model based on the centroid signature in the spatial domain (denoted **S_CSCD**; Section 3.3),
- contour *SRCD* model based on the centroid signature in the frequency domain (denoted **S_SRCD**; Section 3.3).

6. **Center-bias:** Gaussian at the image center (denoted **Center**).

3.5.2 Experimental Results and Discussion

We evaluated the prediction ability of computational models on fixation data from our egocentric dataset to measure the effects of static features on attention (see example saliency maps in Figure 3.25). First, we compared the models using NSS scores from the whole dataset. The results indicate that egocentric attention is most affected by low-level monocular features – intensity, color and orientation. The highest NSS scores were achieved by **GBVS_O** (1.79), **SPX_I** (1.65) and **Itti_O** (1.63).

However, participants' individual NSS scores reported in Table 3.6 indicate significantly different gaze behavior among participants. Therefore, our next analysis relies only on individual saliency scores per frame.

Individuality of Egocentric Attention

To deeper analyze the individuality of egocentric vision, we employed Kruskal-Wallis H tests with Dunn-Bonferroni posthoc comparisons for frame NSS scores of each saliency model (see Figure 3.27, 3.28, 3.29, 3.30, 3.31 and 3.32)[12].

The frequency of statistically significant differences between each participant pair is visualized in Figure 3.26. We found statistically significant differences between participants for each computational models including center-bias. Surprisingly, participant #1 differs from others in 15 computational models which makes him/her the most dissimilar one. Furthermore, fixations of two viewers are significantly closer to their central visual field than the rest of subjects ($\chi^2(5) = 393.9; p < .001$). In addition, we found differences in saccade lengths – two participants have significantly shorter saccades than others ($\chi^2(5) = 380.6; p < .001$).

The revealed differences in viewers' gaze behavior could also indicate that attention in real environments is strongly guided by top-down factors that we ignored in our analysis such as object identification and surprising stimuli over time.

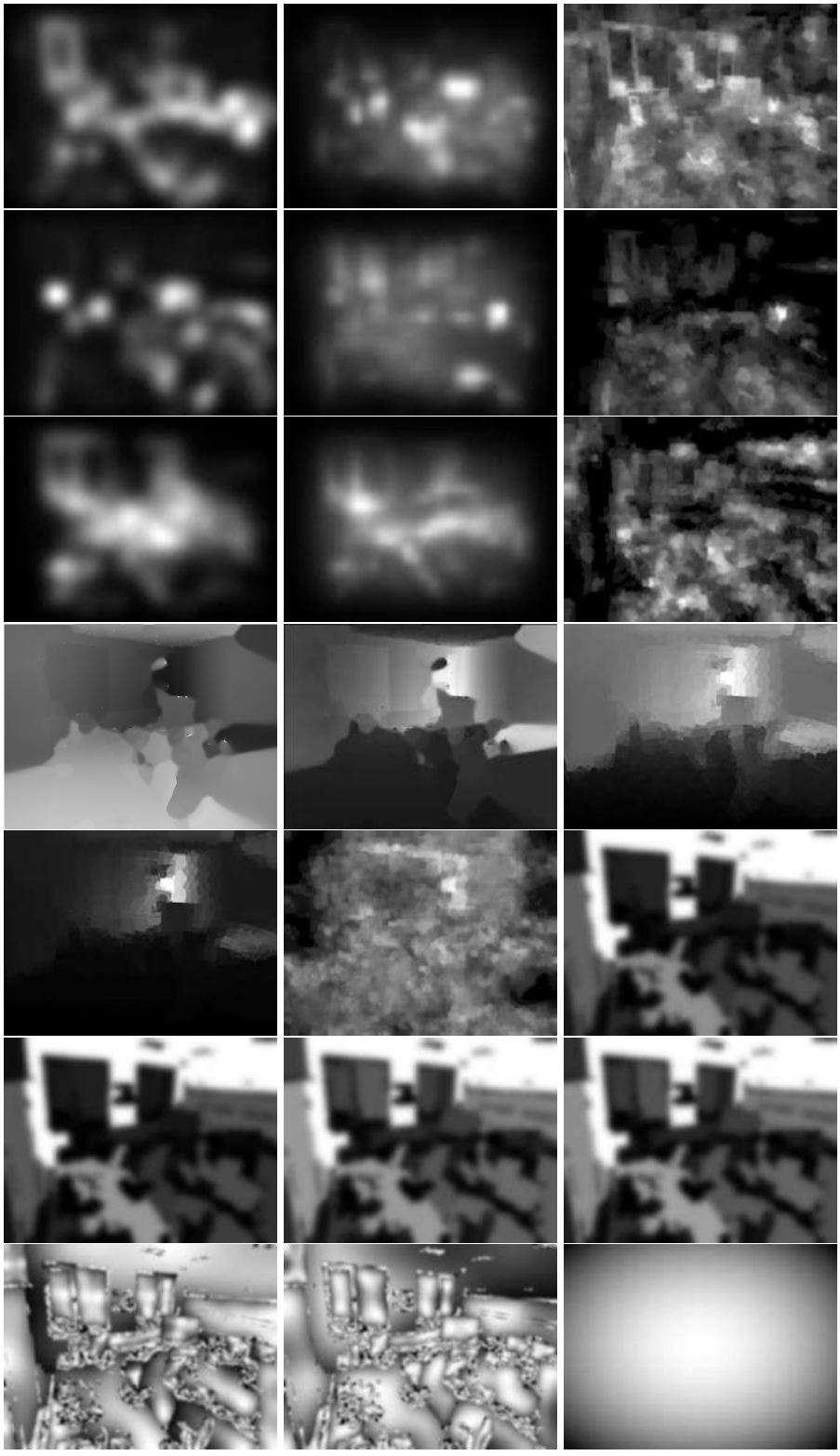


Figure 3.25: Saliency maps for the video frame shown in Figure 3.24(b). From left to right, the 1st row – Itti_I, GBVS_I, SPX_I; the 2nd row – Itti_C, GBVS_C, SPX_C; the 3rd row – Itti_O, GBVS_O, SPX_O; the 4th row – D_linear, D_nonlinear, DC_linear; the 5th row – DC_nonlinear, SPX_D, S_perimeter_intra; the 6th row – S_perimeter_inter, S_eqdiameter_intra, S_eqdiameter_inter; the 7th row – S_CSCD, S_SRCD and Center.

Table 3.6: Individual NSS scores (the highest value is bold; the second and the third highest values are underlined). Results indicate strong differences of attention bias among participants. Low-level monocular features such as intensity, color and orientation are most attentive, in general.

Feature	**Model**	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
Intensity	Itti_I	0.97	1.52	<u>1.79</u>	1.58	1.68	<u>1.77</u>
	GBVS_I	0.99	1.65	<u>1.80</u>	1.62	1.61	1.66
	SPX_I	<u>1.22</u>	<u>1.85</u>	1.78	**1.71**	<u>1.82</u>	1.69
Color	Itti_C	0.99	1.11	1.38	1.11	1.69	1.33
	GBVS_C	1.09	1.39	1.27	1.48	1.85	1.56
	SPX_C	0.73	0.90	1.20	1.48	1.73	1.59
Orientation	Itti_O	1.11	<u>1.87</u>	1.65	<u>1.66</u>	<u>1.85</u>	<u>1.79</u>
	GBVS_O	<u>1.20</u>	**2.00**	**2.02**	<u>1.70</u>	**2.04**	**1.95**
	SPX_O	0.94	1.10	1.27	1.19	1.14	1.51
Depth	D_linear	0.14	-0.10	0.29	0.50	-0.23	0.09
	D_nonlinear	0.45	0.66	0.28	-0.03	0.88	0.52
	DC_linear	0.49	0.30	-0.19	-0.28	0.64	-0.06
	DC_nonlinear	0.60	0.16	-0.29	-0.33	0.56	-0.10
	SPX_D	**1.27**	1.12	1.17	1.27	1.18	1.23
Shape	S_perimeter_intra	0.26	-0.35	-0.51	-0.21	-0.50	-0.44
	S_perimeter_inter	0.30	-0.33	-0.48	-0.15	-0.45	-0.41
	S_eqdiameter_intra	0.14	-0.42	-0.58	-0.36	-0.60	-0.54
	S_eqdiameter_inter	0.18	-0.40	-0.55	-0.29	-0.56	-0.51
	S_CSCD	0.79	0.91	0.89	0.85	0.99	0.80
	S_SRCD	0.86	1.22	1.18	1.06	1.18	1.20
Center		0.98	1.25	0.99	1.08	1.29	1.03

Influence of Static Features on Egocentric Attention

Next, we analyzed feature effects on attention of each participant using frame NSS score. We utilized the repeated measures ANOVA with Bonferroni-adjusted posthoc comparisons to find out which saliency model estimates feature saliency with the highest precision.

As we already pointed out, *simple features such as intensity, color and orientation play an important role in egocentric vision*. Intensity saliency is best modeled by **SPX_I**, with the significantly highest frame NSS scores for 3 individuals (#1: $F(2, 2268) = 31.12; p < .001; \eta^2 = .027$; #2: $F(1.982, 1692.379) = 31.65; p < .001; \eta^2 = .036$; Huynh-Feldt correction; #5: $F(2, 1812) = 15.89; p < .001; \eta^2 = .017$; Figure 3.27). In terms of color saliency, **GBVS_C** outperforms other models, with the significantly highest frame NSS scores for 2 viewers (#2: $F(1.755, 1499.184) = 60.04; p < .001; \eta^2 = .066$; Huynh-Feldt correction; #5: $F(1.897, 1718.282) = 4.855; p = .009; \eta^2 = .005$; Huynh-Feldt correction; Figure 3.28). This indicates that *DKL color space is more suitable to highlight attentive stimuli than red-green and blue-yellow opponent pairs*. Orientation saliency is efficiently estimated by **GBVS_O**, with the significantly highest frame NSS scores for 5 participants (#1: $F(1.298, 1472.240) = 23.24; p < .001; \eta^2 = .020$; Greenhouse-Geisser correction; #2: $F(1.607, 1372.690) = 234.4; p < .001; \eta^2 = .215$; Huynh-Feldt correction; #3: $F(1.355, 875.578) = 76.89; p < .001; \eta^2 = .106$; Greenhouse-Geisser correc-

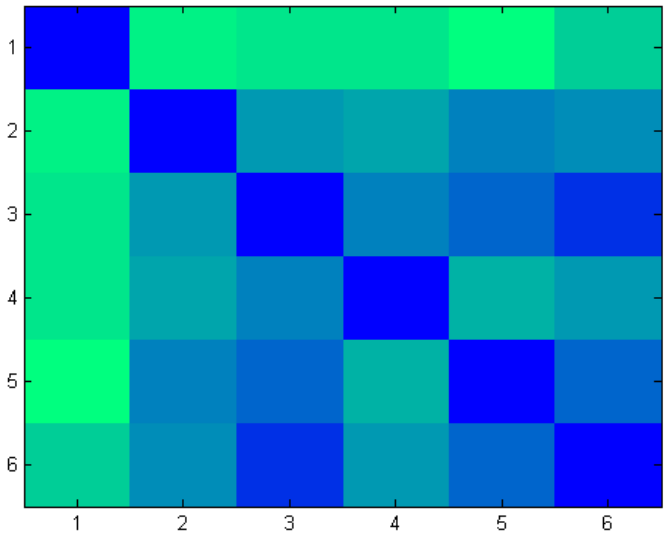


Figure 3.26: Participants' dissimilarity heatmap. The matrix visualizes the number of computational models with statistically significant differences between frame NSS scores of each participant pair. The differences are computed using Kruskal-Wallis H tests followed by Dunn-Bonferoni posthoc tests. The highest dissimilarity is represented by green color.

tion; #5: $F(1.672, 1515.191) = 259.6; p < .001; \eta^2 = .223$; Huynh-Feldt correction; #6: $F(1.389, 1461.471) = 53.60; p < .001; \eta^2 = .048$; Greenhouse-Geisser correction; Figure 3.29).

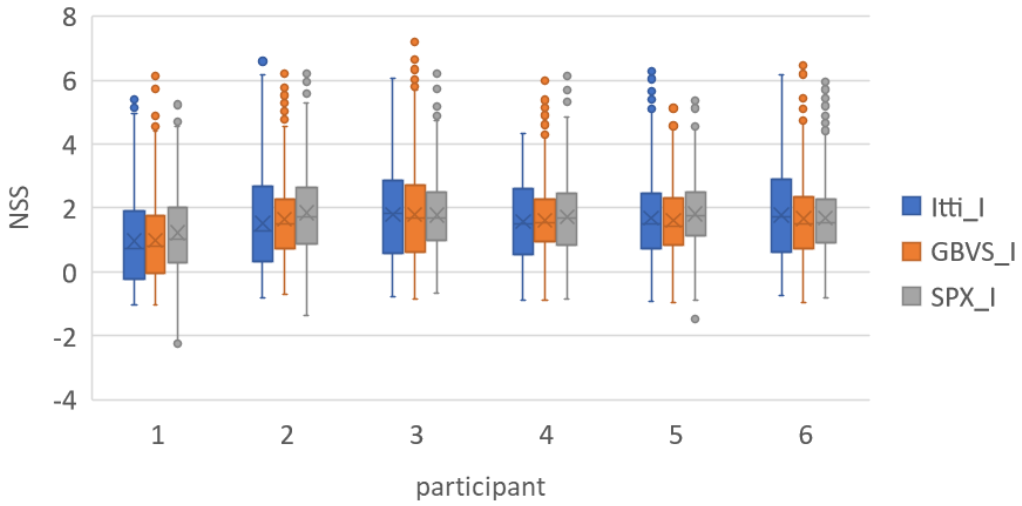


Figure 3.27: Frame NSS scores for intensity saliency.

Object distance (modeled by **D_linear** and **D_nonlinear**) *and global depth contrasts* (modeled by **DC_linear** and **DC_nonlinear**) *have a negligible effect on stereoscopic attention*. The results also showed a large diversity of the depth effects among participants. The poor quality could be partially attributed to missing background subtraction because regions closest to viewers were often floors and walls of the room (see Figure 3.24(b)). Comparing frame NSS scores showed that depth-weighting approach based on our experiment

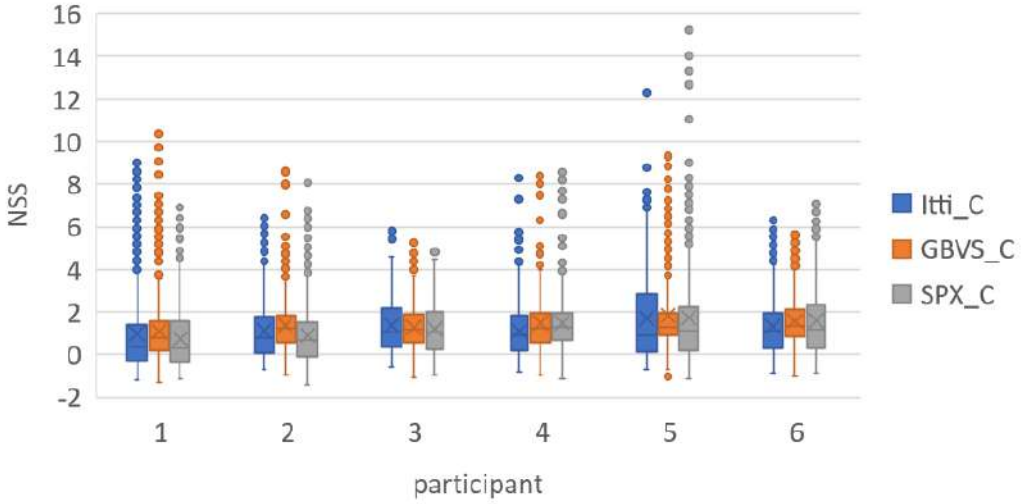


Figure 3.28: Frame NSS scores for color saliency.

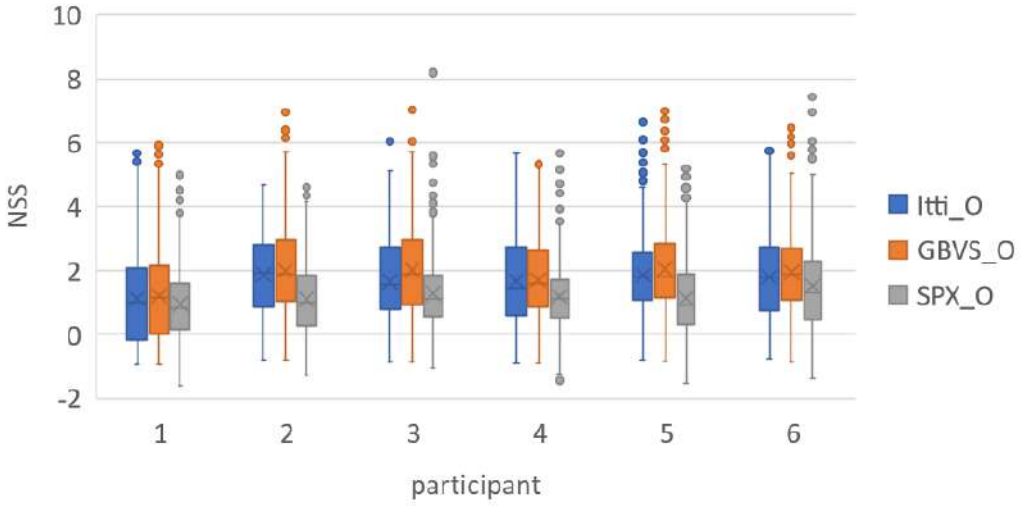


Figure 3.29: Frame NSS scores for orientation saliency.

(**D_nonlinear**; see Section 3.4) significantly improved saliency estimation of standard (linear) depth weighting (**D_linear**) for 4 subjects. In other words, their *attention was often drawn to more distant areas of the room*. However, the worsen performance of the other 2 participants indicates the need of more controlled experiment of depth saliency with larger depth range, in the future. On the other hand, our experimental method for global contrast (**DC_nonlinear**) brought a significant improvement in performance only for one participant in comparison with a standard contrast model (**DC_linear**). Furthermore, the results highlight *the importance of local depth contrasts in egocentric vision* estimated by **SPX_D**. This model clearly outperforms other depth models for all individuals (#1: $F(1.742, 1974.930) = 152.5; p < .001; \eta^2 = .119$; Greenhouse-Geisser correction; #2: $F(1.769, 1510.519) = 247.5; p < .001; \eta^2 = .225$; Greenhouse-Geisser correction; #3: $F(1.807, 1167.520) = 347.5; p < .001; \eta^2 = .350$; Greenhouse-Geisser correction; #4: $F(1.827, 1045.108) = 471.0; p < .001; \eta^2 = .452$; Greenhouse-Geisser correction; #5: $F(1.549, 1402.963) = 289.7; p < .001; \eta^2 = .242$; Greenhouse-Geisser correction; #6: $F(1.583, 1665.158) = 406.3; p < .001; \eta^2 = .279$; Greenhouse-Geisser correction; Fig-

ure 3.30).

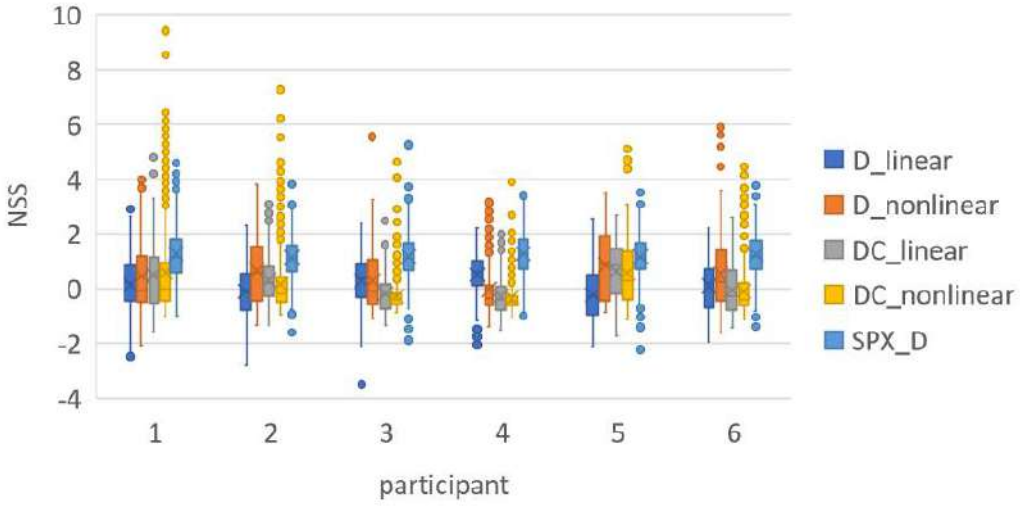


Figure 3.30: Frame NSS scores for depth saliency.

The negative NSS values of **S_perimeter_intra**, **S_perimeter_inter**, **S_eqdiameter_intra** and **S_eqdiameter_inter** indicate that *fixations are rarely located in the largest regions* in the room. This poor saliency estimation is partially caused by missing background subtraction because floors and walls were usually the largest uniform areas of individuals' view. *In contrast to intra- and inter-shape saliency models, contour saliency models* (**S_CSCD** and **S_SRCD**) which represent objects by the centroid distance signature *predict attentive shapes with the highest precision* for all subjects (#1: $F(1.546, 1752.712) = 202.4; p < .001; \eta^2 = .151$; Greenhouse-Geisser correction; #2: $F(1.318, 1125.708) = 1130; p < .001; \eta^2 = .570$; Greenhouse-Geisser correction; #3: $F(1.376, 889.190) = 1010; p < .001; \eta^2 = .610$; Greenhouse-Geisser correction; #4: $F(1.441, 824.105) = 561.6; p < .001; \eta^2 = .495$; Greenhouse-Geisser correction; #5: $F(1.384, 1254.128) = 1788; p < .001; \eta^2 = .664$; Greenhouse-Geisser correction; #6: $F(1.421, 1495.199) = 1391; p < .001; \eta^2 = .569$; Greenhouse-Geisser correction; Figure 3.31). This finding is consistent with our previous shape experiment where our **S_SRCD** model achieved leading performance (see Section 3.3).

Finally, our experiment confirms a strong center bias of egocentric gaze behavior (see Figure 3.32).

We further checked the effects of salient features on attention over time. We found only moderately positive correlations with inter- and intra-shape saliency models for 2 individuals. Moderately negative correlations with intensity, color and orientation were observed only for a single individual. However, while the length of room exploration did not strongly correlate with salient features, we found much stronger correlations with feature saliency in first seconds of the experiment, e.g. the participant shown in Figure 3.33. This suggests that simple low-level cues predominantly capture attention in the initial perception and then complex features become more and more dominant.

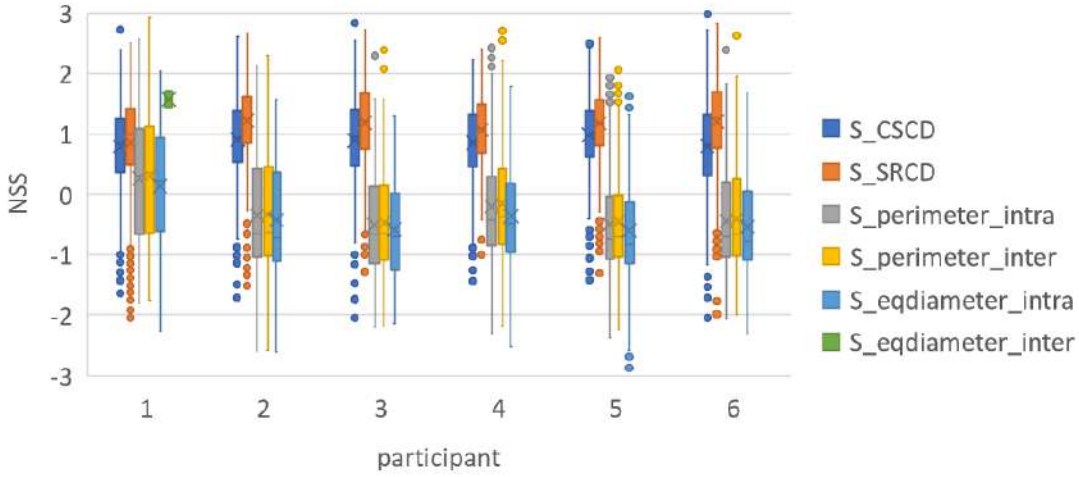


Figure 3.31: Frame NSS scores for shape saliency.

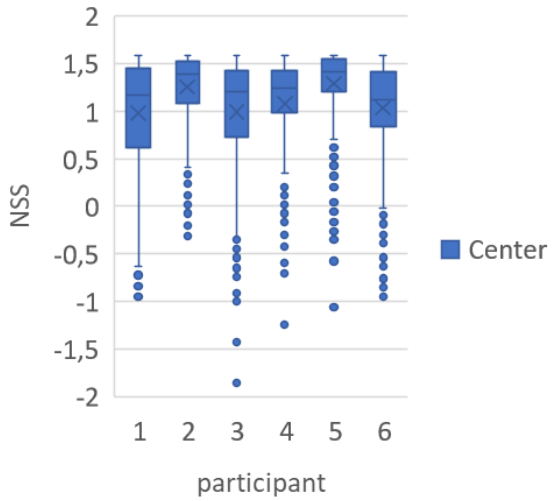


Figure 3.32: Frame NSS scores for center-bias.

3.5.3 Summary

Our experiment explored egocentric attention affected by static salient stimuli in a real indoor environment including depth saliency, in contrast to previous studies. Comparing participants' fixations revealed significant differences in human gaze behavior. We also found out that attention is primarily influenced by simple feature contrasts as intensity, color and orientation with a strong center bias.

Furthermore, we showed that fixations of most viewers are more concentrated in farther locations than the closer ones. This confirms the conclusion of our depth experiment (see Section 3.4) that egocentric stereoscopic attention differs from image viewing conditions and therefore a specialized saliency model is needed. In addition, the results revealed that local contrasts are significantly the most important factor of depth saliency.

We confirmed the conclusion of our shape experiment too (see Section 3.3). Salient contour parts dominate over saliency resulted from global shape characteristics and contrasts.

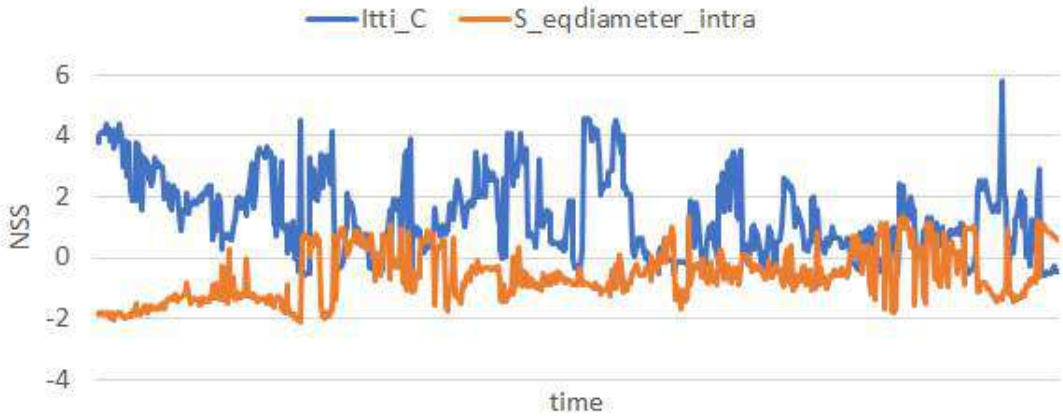


Figure 3.33: Relationship between frame NSS values and exploration time of a participant. There is a moderately negative correlation with color predicted by **Itti_C** ($r = -0.38; p < .001$) and a moderately positive correlation with object size predicted by **S_eqdiamter_intra** ($r = 0.410; p < .001$). However, when we look into the first video frames of the experiment, we found very strong correlations with the features (first 100 frames – **Itti_C**: $r = -0.855; p < .001$; **S_eqdiamter_intra**: $r = 0.776; p < .001$). In other words, while color contrasts become less attentive, there are more fixations in larger objects as time passes.

The results also indicate that the above mentioned effects did not remain constant over time. Simple low-level features seem to affect attention more rapidly, whereas complex features such as shape have a delayed effect.

Our experiment was focused solely on static low-level features. In the future, it will be important to analyze high-level features too, which could be a reason for variance in human gazes. Future work should consider to build and explore larger datasets with more participants that employ both static and dynamic stimuli.

The weak point of actual egocentric studies is evaluation of computational saliency models too. We therefore suggest to register individual's views to build their 3D fixation heatmaps over the whole exploratory space instead of evaluation based only on a single fixation per frame.

3.6 Emotionally-Tuned Visual Attention

While psychological studies have confirmed a connection between emotional stimuli and visual attention, there is a lack of evidence, how much influence individual's mood has on visual information processing of emotionally neutral stimuli. In contrast to prior studies, we explored if bottom-up low-level saliency could be affected by positive mood. We therefore induced positive or neutral emotions in 10 subjects using autobiographical memories. We recorded eye-tracking data which are publicly available under free-viewing and task-specific conditions. We discuss differences in human gaze behavior between both emotions and relate their fixations with bottom-up saliency predicted by a traditional computational model [96].

3.6.1 Motivation

Emotion is capable to influence our visual perception on many levels [220]. Affective states can enhance or worsen individual's selective attention [62]. Maekawa et al. [145] found a positive effect of happiness on performance in serial visual search. On the other hand, Most et al. [158] suggested that positively arousing stimuli can distract attention.

Saliency models are commonly used to predict human fixations in natural images for emotionally neutral conditions. While recent saliency models aimed to predict attention for affective stimuli [137, 46], conventional saliency models have been never evaluated under a particular emotional state of observers.

3.6.2 Emotion-Based Visual Attention Experiment

We can assume from Maekawa et al.'s experiment [145] that positive mood is associated with faster serial search. We extended their investigation of emotions to different scenes and different task types including free viewing and visual search tasks. The primary goal of this experiment is to assess the effect of positive emotional state induced by imagination on visual attention, particularly bottom-up saliency and search efficiency, compared to neutral mood manipulation. The whole dataset is publicly available[20].

Hypotheses

Our eye-tracking experiment is based on the broaden-and-build theory [62] so positive emotions improve user's attention and cognition:

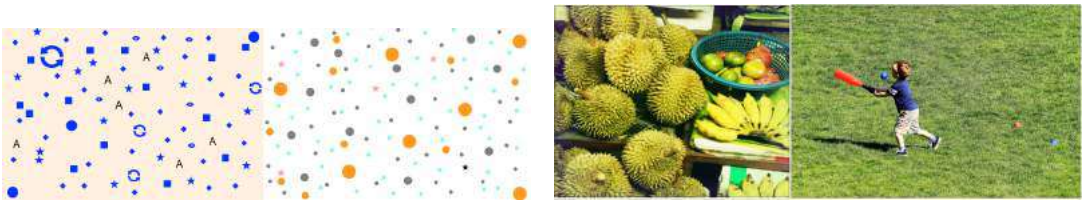
- **H1:** *Positive emotion enhances visual performance.* We assume that the broadening effect of positive affective state is beneficial for visual search. It could motivate users to efficiently complete tasks, similarly to Maekawa et al.'s study [145]. We therefore expect that a target is detected faster when experiencing a positive mood.
- **H2:** *Attention of users experiencing a positive emotion is strongly guided by bottom-up factors.* Positive affective state means a broader spatial focus of attention, thus we expect that it produces stronger effects of saliency. We therefore hypothesize as follows:
  - *H2.1:* Regions fixated in a positive state differ from a neutral state while fixations of users experiencing the same emotion is more consistent.
  - *H2.2:* Similarity between bottom-up saliency maps and human fixations is higher for users affected by a positive mood than a neutral mood.

Image Data

We investigated visual attention using 38 displays with valence-neutral stimuli that are suitable for target search tasks. Participants were shown our own 24 simple images as well as 14 more complex, natural images selected from the COCO dataset [135], as shown in Figure 3.34. Simple images contains multiple non-overlapping objects that are on backgrounds

[20] `https://vgg.fiit.stuba.sk/2018-08/emotions-attention/`

of uniform color. We can thereby expect that they affect users' attention more strongly in bottom-up manner than natural images.



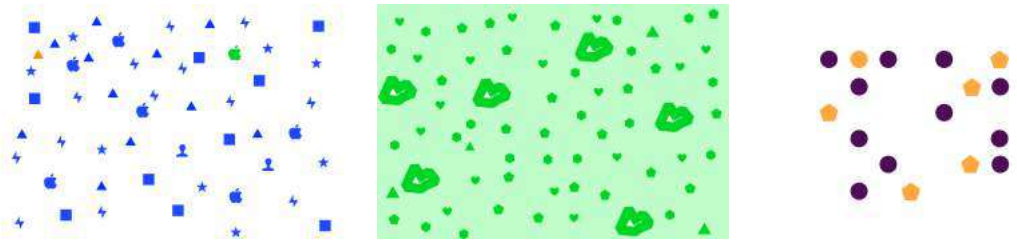
(a) Simple images. (b) Natural images.

Figure 3.34: Examples of images used in the experiment.

Tasks

We formulated an easily solvable task for each image that falls into one of the following categories (see examples in Figure 3.35):

1. **FO-task:** *find* a *target object* among non-targets based on specific criteria (14 tasks),
2. **FAO-task:** *find any* of *target objects* among non-targets that meets specific criteria (5 tasks),
3. **FU-task:** *find* a *unique object* whose visual attributes differ from other objects' attributes (9 tasks),
4. **V-task:** *free view* of an image (5 tasks),
5. **M-task:** *memorize* the image content (5 tasks).



(a) Find an orange triangle. (b) Find a triangle. (c) Find a unique object.

Figure 3.35: Examples of FO-task, FOA-task and FU-task, respectively. To complete a FO-task, subjects had to select the only object that has an orange color and a triangular shape. In contrast, four objects satisfy the target description in a FOA-task, but participants had to detect only one of them. A specific description of a target is missing in a FU-task. The goal of this task is to search for a unique object – the only one that is orange and circular at once.

A target object in simple images differs from non-targets in up to 3 visual attributes, such as color, shape, orientation or a combination of them. The number of targets in FOA-tasks varies from 2 up to 4. In contrast to the FU-task, the FO-task and the FOA-task describe visual attributes of a target that should be searched. Such a target description is not included in the FU-task. The aim of this task is to select the object whose combination of visual attributes is unique in a given image.

No time limit was set for FO-tasks, FOA-tasks and FU-tasks. For V-tasks and M-tasks, images were displayed for a 4-second period.

We expect that fixations of users are strongly guided by visual search tasks, thereby less by bottom-up factors compared to exploration in V-tasks and M-tasks.

Mood induction and measures

Participants in our experiment were randomly induced to experience either a positive mood or a neutral mood by recalling own personal memories, similarly to the experiment conducted by Becker and Leinenger [9]. The mood induction took place in a small laboratory with an experimenter.

First, participants were asked to rate their actual mood in the *I-PANAS-SF* [205, 109][21].

To induce a positive mood, participants were asked to recall a very happy event from their own lives. They were given a minute to imagine this event as vividly and as concretely as possible. Then, following the study of Watkins and Moberly [204], they answered questions regarding this happy memory (e.g.“Who and what can you see, feel and hear?", “How does the event unfold moment-by-moment?”).

On the other hand, subjects in the neutral mood induction were asked to recall the route they took to arrive at the place of experiment for a 1-minute period. Then, they wrote down as many details about the route as possible[22].

After the mood induction, participants ranked their current affective state again as a manipulation check. Finally, participants in the positive mood were told to recall their happy memory again.

Participants

Participants who attended our experiment were 10 members of academical staff and students (6 females). Participation in the experiment was voluntary. Each participant gave written informed consent to the study and received information regarding the experiment. Age of participants ranged from 19 to 36 years with the mean age of 23.9 years (SD = 5.28). Half of them were assigned to the positive condition and the other half to the neutral condition.

Experimental Procedure

The whole experiment was realized in two parts. First, participants were induced to experience either a positive mood or a neutral mood in a laboratory with an experimenter. Then, they were instructed to go to the nearby laboratory where the eye-tracking experiment took place. The first and the second part of the experiment took about 20 minutes and 5 minutes, respectively.

[21] In the International Positive and Negative Affect Schedule Short Form (I-PANAS-SF), users rate 10 terms – 5 terms measuring positive affect (active, determined, attentive, inspired and alert) and 5 terms measuring negative affect (afraid, nervous, upset, hostile and ashamed) using a 5-scale range that ranges from very slightly or not at all to extremely [205, 109].

[22] We assumed that description of the route to the lab is associated with non-emotional stimuli, thereby reducing their happiness level, compared to the positive group.

In the second part, participants were shown the same set of images on a computer screen, grouped by the type of tasks they had to perform. We randomized the order of these groups as well as the order of images in each group. Each group started with an example image to familiarize with the task type.

Participants were asked to find a target as quickly as possible (FO-task, FOA-task and FU-task). As soon as they localized the target, they pressed the space bar.

Measures and analysis of visual attention

We recorded participants' fixations and task completion time.

The *task completion time* (TCT) is a period of time from the initial display of a given image to press of the space bar. This measure is used to test if positive mood makes the target search faster.

Furthermore, we measured the *fixation similarity* between participants and their blurred fixation maps. For each image, we defined the *intra-emotion similarity* as the average correlation between each pair of participants' fixation maps recorded in the same affective state. The *inter-emotion similarity* corresponds to the average similarity between fixation maps that differ in the affective state. These measures should reveal whether current emotions affect where to direct the gaze.

Finally, we estimated bottom-up saliency by Itti et al.'s computational model [96] to assess this effect during both affective states. Saliency maps are evaluated using AUC and NSS (see definitions of evaluation scores in Section 2.10).

Apparatus

The gaze data were recorded by Tobii X2-60 eye-tracker at 60 Hz and processed by Tobii I-VT fixation filter. Images were displayed on a 24.1-inch computer screen with a resolution of 1920×1080 pixels at a viewing distance of approximately 60 cm.

3.6.3 Experimental Results

First, we checked if the mood manipulation was effective, as can be seen in Figure 3.7 (participants with prefix "N" and "P" were assigned to neutral and the positive conditions, respectively). Observers before mood induction had a mean positive-affect score of 16.7 (SD = 3.53)[23]. The positive manipulation changed the affective rating only slightly in comparison to the neutral group, though. After mood induction, neutral subjects had the mean score of 17.2 (SD = 2.78), while the score of positive subjects increased to 17.8 (SD = 4.92). However, we found significant differences among participants' rankings (see the score of subject P5).

[23]Positive-affect score refers to the ranking sum of PANAS items measuring positive affect.

Table 3.7: Positive-affect values before and after the mood induction procedure (N – neutral, P – positive).

Manipulation	**N1**	**N2**	**N3**	**N4**	**N5**	**P1**	**P2**	**P3**	**P4**	**P5**
Before	17	18	15	17	17	19	16	14	21	8
After	14	18	15	18	21	18	22	17	22	10

Visual search performance

To test the effect of emotions on visual performance, we measured task completion time (TCT) of target search tasks (FO, FOA, FU), as listed in Table 3.8. In general, tasks with explicit description of target appearance (FO and FOA) are solved considerably faster than FU-tasks.

The results showed that *positive mood does not reduce time to complete a task, but could even worsen the efficiency*. This negative influence is significant in the FU-task, in which individuals needed 3 more seconds on average to find a target compared to those under neutral condition. This finding indicates that positive mood acts as a distractor what contrasts with the broadening effect formulated in **H1**, thereby rejecting the hypothesis.

Table 3.8: The average task completion time (sec) of target search tasks (N – neutral, P – positive).

Task	**N1**	**N2**	**N3**	**N4**	**N5**	**P1**	**P2**	**P3**	**P4**	**P5**
FO	2.60	2.10	1.42	3.26	1.94	3.32	5.96	2.86	3.32	1.59
FOA	1.28	1.24	1.21	0.78	1.34	1.35	1.90	1.73	1.11	0.83
FU	8.44	6.74	4.55	4.24	6.79	11.83	7.96	6.20	7.37	12.94

Fixation similarity

To explore the difference between regions that attracted participants' attention, we measured intra- and inter-emotion similarities (see Figure 3.36). In general, searching for specifically defined targets leads to the lowest fixation similarity for both affective states. On the other hand, memorizing the image contents affects participants more similarly than other tasks.

We also observed that fixations of the positive group are less correlated for the M-task than those of the neutral group. For the other tasks, the affective states seem to have a negligible influence on the fixation similarity. *In other words, an emotional state does not result in coherent fixations between users*, thereby rejecting hypothesis **H2.1**.

Fixation-saliency similarity

Finally, we tested a relationship between fixations and bottom-up saliency. We therefore computed saliency maps by Itti et al.'s model [96] (see examples in Figure 3.37 and 3.38) and compared with participants' fixations using AUC and NSS scores, as shown in Figure 3.39.

According to the average AUC scores, saliency is not related to emotions. Looking into the average NSS scores, we found differences for the FU-task and exploration tasks (M and V).

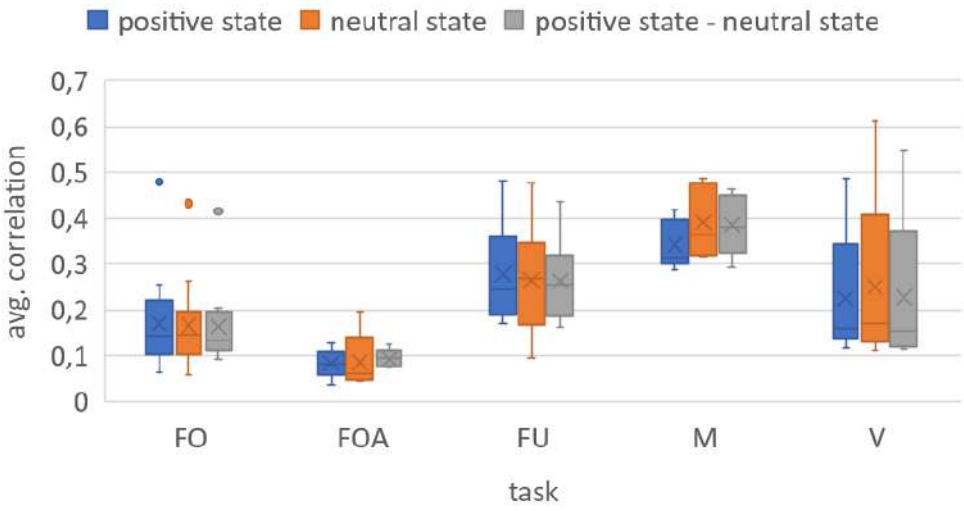
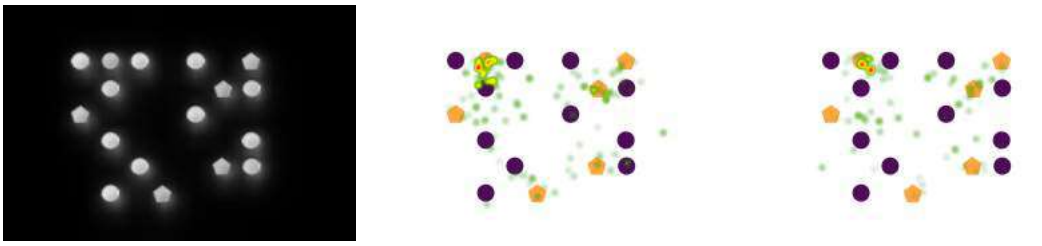


Figure 3.36: Intra- and inter-emotion similarities of tasks which subjects solved.



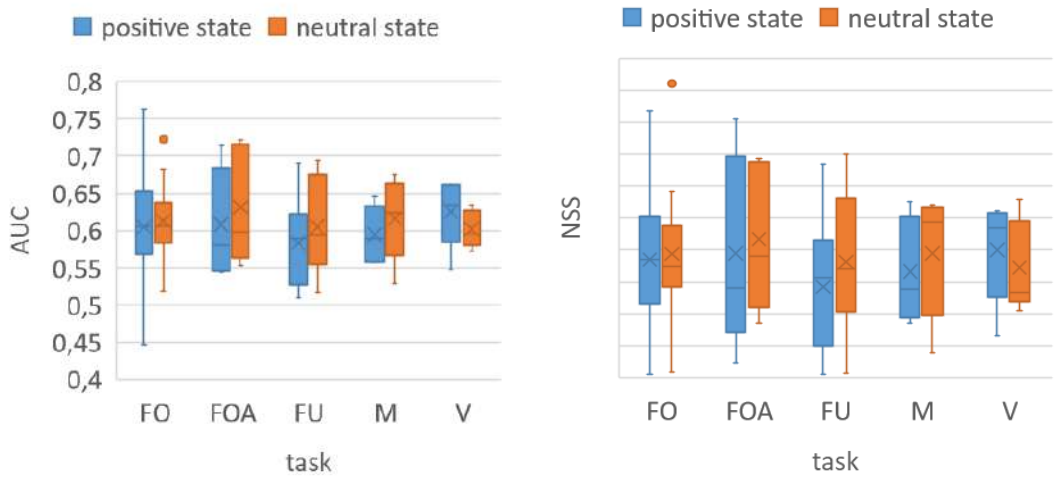
(a) Saliency map. (b) Positive emotion. (c) Neutral emotion.

Figure 3.37: Saliency map [96] and fixations during the FU-task for the image in Figure 3.35 (right).

(a) Saliency map. (b) Positive emotion. (c) Neutral emotion.

Figure 3.38: Saliency map [96] and fixations during free-viewing of a natural image.

For the FU-task and the M-task, however, the saliency model achieved the higher score for participants experiencing a neutral mood. On the other hand, the score of the positive group was significantly better for the V-task. This means that bottom-up saliency has a stronger effect on attention after the positive mood induction only during task-free exploration. We therefore also have to reject hypothesis **H2.2**.



(a) AUC. (b) NSS.

Figure 3.39: Evaluation scores for Itti et al.'s saliency [96].

3.6.4 Discussion

Our research aimed to study the effect of positive induced mood on visual attention and visual search performance for emotionally neutral stimuli. Though the mood induction affected only slightly how participants assessed their affective state, we found some differences between both groups of subjects.

We assumed that experiences of positive emotions broaden individual's attention, as proposed by the broaden-and-built theory [63, 62]. Therefore, we expected fast task completions and strong bottom-up effects on attention of the positive group of participants.

However, in contrast to the finding of Maekawa et al. [145], the results showed that the recall of happy memories is not associated with considerably faster target search, but it could even distract individuals from their focus (rejecting **H1**). A similar effect was observed in the Most et al.'s experiment [158] for negative and positive arousing stimuli. A possible explanation for our finding could be that attention was turned inwards to own inducted autobiographical memory rather than to the target stimulus.

We also expected that inducted emotion could affect individuals where to look. While we did not observe significantly similar fixation patterns for participants within as well as between induced emotion (rejecting **H2.1**), the results suggest the interaction between induced mood and bottom-up saliency [96]. However, positive emotions broadened the influence of saliency only in task-free viewing conditions. When participants were given a particular task, this effect either disappeared or was even reversed (rejecting **H2.2**). This finding might be explained by individual's different engagement levels that could affect mood and attention too [186, 162, 99]. Smilek et al. [186] suggested that passive task instructions allow observers to hand off attentional control to automatic (bottom-up) processing. Similarly, when following our explicit instructions, interest in a particular task is assumed to be higher than free image exploration and the effect of salient features is presumably reduced. In addition, cognitive effort to successfully and quickly complete visual search and memory tasks could suppress positive mood induction and thereby also narrow attention, compared to a more relaxed task-free condition. We therefore reject hypothesis **H2** and suggest that

broadening bottom-up attention by positive mood induction relates to task difficulty and task engagement.

3.6.5 Summary

Our experiment confirms that individual's emotional state could participate in visual attention mechanisms. Unlike the majority of studies which have examined visual attention in stimuli of high valence, we induced subjects into positive and negative mood and studied their attention in non-emotional images when performing different tasks. Even though most participants assessed their affective states similarly, we observed behavioral changes between both mood induction procedures.

In contrast to the broaden-and-built theory [62, 63], we found that positive induction could even impair visual performance in search tasks. On the other hand, we observed a higher diversity of fixations regardless of a mood we induced to participants. However, we found that attention bias towards bottom-up features varies across task type. While saliency under positive conditions is stronger when freely exploring images, bottom-up attention becomes weaker when solving tasks, compared to neutral conditions. We therefore speculate that broadening attention in terms of bottom-up processing might be associated with a low level of engagement in the task.

In future work, it will be important to involve more participants and emotions in experiments. In addition, we should verify if the distracting effect can be also observed after other mood induction procedures. We suspect that the level of subjects' task engagement should be more controlled in future to fully verify the broadening effect of emotions on attention.

3.7 Visual Attention during Task-Based Analysis of Information Visualizations

The way users observe a visualization is affected by salient stimuli in a scene as well as by domain knowledge, interest, and tasks. While recent saliency models manage to predict users' visual attention in visualizations during exploratory analysis, there is little evidence how much influence bottom-up saliency has on task-based visual analysis. Therefore, this section presents our own eye-tracking study whose aim is to determine user's gaze behavior when solving three low-level analytical tasks using charts from the MASSVIS database [21]. To the best of our knowledge, such an experiment has not been performed so far. Hence, we made fixation data from this experiment publicly available. We also compared our task-based eye tracking data to the data from the original memorability experiment by Borkin et al. [20]. Our experiment was already published in [236].

3.7.1 Motivation

Eye tracking is becoming a popular alternative for evaluating visualizations, compared to classic completion time and accuracy evaluations [124]. Mostly, it is employed to understand how users read a particular encoding, such as parallel coordinates [184], or to compare visual exploration of different encodings, such as different tree layouts [27, 26], graph layouts

[171], linear and radial charts [72], or user strategies for sorting in tabular visualizations [113]. Task-dependent areas of interest can be used to assess in which order and frequency users fixate crucial chart elements when decoding the visualization [72, 73, 175]. Besides, visual saliency predicted by computational models is applied in visualization research. Haass et al. [79] compared the performance of three different saliency models between the cat2000 dataset [14] and the MASSVIS dataset from Borkin et al.'s memorability experiment [20] using eight different comparison metrics. They found that saliency models performed worse for information visualizations than for the natural images. One possible explanation by the authors is that text labels in visualizations attract the users' attention to a higher extent than indicated by the saliency models. Indeed, Matzen et al. [153] and Bylinskii et al. [30] showed that most fixations in visualizations can be accounted to regions containing text. In the DVS model [152], Matzen et al. linearly combined a variation of Itti et al.'s model [96] with text saliency and successfully outperforms standard saliency models. While it has been shown that bottom-up factors captured by this data visualization model have a strong influence on users' visual attention during exploratory visual analysis, it is still unknown how strong top-down guidance influences attention during task-based visual analysis.

3.7.2 Memorability Experiment by Borkin et al.

In the memorability experiment conducted by Borkin et al. [20], eye tracking data was gathered in two separate treatments using various charts of the MASSVIS dataset [20]: In a 10-second encoding phase, users examined about 100 visualizations without a pre-defined goal. In the subsequent 2-second recognition phase, they were asked to indicate whether they had seen the particular visualization before. Finally, in the 20-minute recall phase without eye tracking, they were presented with small blurred versions of recognized visualizations, and were asked to write down everything they remember being shown.

The outcomes of this experiment give indications which visualization elements attract the users' attention, and which elements make a visualization memorable. Human recognizable objects, such as photographs and pictograms enhance memorability of visualizations. Data redundancy also improves the ability to recognize visualizations. Text elements, particularly titles, were most fixated regions in the encoding phase and also most frequently mentioned elements in the recall phase.

3.7.3 Task-Based Visual Analysis Experiment

We can assume that the impact of top-down factors on user's visual attention varies for the activities with visualizations. Exploratory analysis is less driven by top-down factors than confirmatory analysis to answer a specific question. For presenting known facts, often highlighting is used to effectively draw the attention to specific regions, but the task-imposed guidance is also low.

In our experiment, we compare confirmatory (or task-based) visual analysis and exploratory visual analysis. The purpose is to test if task-based visual analysis is indeed strongly driven by top-down factors, so that bottom-up saliency has negligible influence on the user's attention during the task.

Hypotheses

H1: *Overall, top-down factors, such as a particular task, play such an important role in guiding visual attention that bottom-up factors have a negligible effect on the recorded fixation patterns.* We reason that fixations of users will be strongly guided by the task during task-based visual analysis. To solve a task, users have to look at pre-defined areas of interest within the visualization, which will require most of their attention. On the other hand, we expect that during exploratory analysis, users will be more strongly guided by bottom-up factors. We therefore expect the following results:

- **H1.1**: Fixations between users solving the same low-level analytical task will be more coherent than when exploring a visualization without a specific task.
- **H1.2**: When solving a low-level analytical task, users fixate on a sequence of specific chart areas in a task-dependent order.
- **H1.3**: The similarity between the recorded fixation maps and bottom-up saliency maps will be higher when users explore a visualization without a specific task than when performing a low-level analytical task.

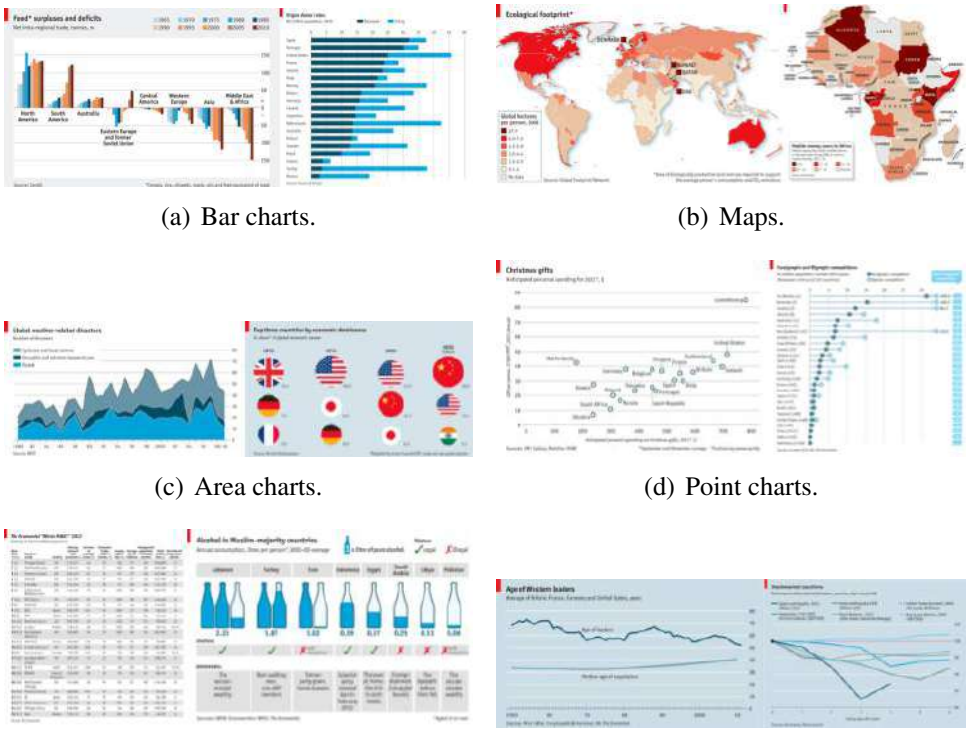
H2: *Bottom-up factors have an influence on visual attention when performing a visual search for a target.* While we assume that bottom-up saliency does not have a strong influence on users' fixations (see H1.3), we do believe that it has an influence on visual search efficiency for target areas when solving a low-level analytical task. In particular, extreme values should stand out in their associated visual channel, for instance as the longest bar in a bar chart or the darkest region in a choropleth map. We therefore assume that extreme values should also show up as salient regions in the saliency maps, and that salient target data points are therefore fixated more quickly than non-salient data points. As a consequence, we assume that users can find extreme values more efficiently than retrieving values of specific items or items associated with specific values. Our specific hypotheses are the following:

- **H2.1**: Efficiency of visual search for a target area depends on the area's visual saliency.
- **H2.2**: Extreme data points show up as highly salient graphical marks in saliency maps.
- **H2.3**: Extreme values can be found most efficiently.

Image Data

Since our hypotheses are not targeted towards a specific visualization type, we chose the MASSVIS database [21] as source for our image data. This database contains around 5000 static visualizations obtained from different online sources, such as news media, blogs, scientific publications or government reports. The contained visualizations are targeted towards a broad audience and are therefore a popular choice to evaluate how non-experts read visualizations [21, 20, 79, 152, 29]. We selected a subset of 30 visualizations from the dataset with the goal to cover a large variety of visualization types, such as bar charts, maps, area charts, tables, point charts and line charts (Figure 3.40) from the "news media" and "infographics" categories. Thereby, we only chose visualizations with associated eye tracking data.

16 charts contain human recognizable objects (HROs) such as pictograms and real objects. We selected visualizations with a rather low average data-ink ratio (ratio of data and non-data elements) of 1.5 (measured on a scale from 1=low to 3=high).



(a) Bar charts. (b) Maps.

(c) Area charts. (d) Point charts.

(e) Tables. (f) Line charts.

Figure 3.40: Categories of visualizations used in the experiment.

The dataset is accompanied by manually defined labels summarized in Table 3.9.

Table 3.9: Original labels of the MASSVIS dataset [21].

Category	**Description**
Annotation	visual elements annotating the data
Axis	axes including tick marks and values
Data	area where data are plotted
Data (type)	visual representation of the actual data
Graphical element	elements not related to data
Legend	legends or keys
Object	HROs
Text	textual elements

Tasks

Visualization research has produced various task taxonomies that formalize activities with visualizations. We adopted the taxonomy proposed by Amar et al. [6] that specifies low-level analytic tasks in the field of information visualization. According to the authors, these are simple, easily solvable tasks, where users have to analyze the visualized data points. In contrast to high-level tasks, users do not require higher-level domain knowledge. Amar et al.'s collection of simple tasks is one of the most frequently cited task taxonomies in the visualization community [22]. We picked the three simplest tasks out of this taxonomy, which are easily solvable across a wide range of visualization types – in particular the visualization types in the MASSVIS database.

For each visualization, we formulated a question according to the task (see examples in Figure 3.41):

- *retrieve value* of a specific data element (RV-task),
- *filter* data elements based on specific criteria (F-task),
- *find* an *extremum* attribute value within a dataset (FE-task).

(a) What is the attendance of Universal Studios Hollywood?

(b) Which German states have an unemployment rate of more than 12%?

(c) In which country do people anticipate to spend the least money for personal Christmas gifts?

Figure 3.41: Target-dependent AOIs of sub-parts of visualizations for the RV-task (a), the F-task (b), and the FE-task (c), respectively. Red, green and blue outlines define the target data points, their item labels and value labels respectively. To complete the RV-task, a target item label has to be searched. Then, the value label of the target data point (i.e., bar) is read. For the F-task, participants search the value label that satisfies the given condition. Then, they search data points (i.e. states) with the color that corresponds to this value and read their names. The scatterplot has data points sorted by the anticipated personal spending on Christmas gifts. Thus, participants only have to find the left-most dot without reading the actual value label.

For comparing task-based visual analysis to more exploratory analysis, we additionally analyzed the eye tracking data of the memorability experiment [20] (Mem-task). For the evaluation, we used fixations from the first phase (encoding phase) of the memorability experiment, when participants were shown visualizations for the first time.

Areas of Interest

To be able to more specifically analyze eye-tracking data with respect to the given task, we additionally defined task-dependent areas of interest (AOIs) for each visualization and low-level task (RV, F, FE), respectively, listed in Table 3.10. They comprise all elements of the visualization that need to be attended to correctly answer the question. It is important to note that not all visualizations contain all AOIs, such as legends. Figure 3.41 shows exemplary visualizations with task-dependent AOIs.

Depending on the task, there is an optimal viewing sequence in which these task-dependent AOIs should be examined in order to answer the question. For their eye tracking experiments, Goldberg and Helfman [73] defined AOIs for three sequential steps required to retrieve values in linear or radial graphs: “find dimension”, “find associated datapoint”, “get data point value”. We adopted these three steps for the three low-level tasks in our experiment (see Table 3.11).

Table 3.10: Task-dependent AOIs.

Category	Description
Value label	textual value label of a target attribute
Value annotation	textual elements annotating values of an attribute
Value legend	legend or keys of attribute values of data points
Data point	target data point
Item label	textual identification of a target
Item legend	legend of item encodings

Table 3.11: Optimal viewing order of task-dependent AOIs.

Task	Step 1	Step 2	Step 3
RV	search item label	map to the item	read the value label
F	search value label(s)	map to the item(s)	read the item label(s)
FE	search value label(s)	map to the item(s)	read the item label(s)
	search item(s)		read the item label(s)

Experimental Design and Procedure

Using a within-subjects design, participants were shown the same subset of 30 visualizations without any repetitions. The order of appearance was counterbalanced with a Latin square across participants. We formulated one RV-task, one F-task and one FE-task for each visualization, but participants solved only one task type per visualization. The order of the types was randomized with the equal distribution of each type and balanced across participants.

Participants had to correctly solve the task as quickly as possible. The procedure of task completion consisted of three steps that were repeated for each visualization:

1. *Task description*: First, participants were shown a question. After they understood and remembered the question, they pressed the spacebar.
2. *Visualization*: Next, they saw a visualization which they should analyze to answer the question. We did not show a central fixation cross before displaying the visualization. In order to keep the same viewing conditions as in the original memorability experiment [20], the task description that would affect participants' scanning sequence, was not displayed in this step of the experiment. As soon as they found the answer, they pressed the spacebar again.
3. *Answer form*: Finally, participants were shown a form where they entered their answer.

The experiment started with three example tasks to familiarize with each task type. The whole experiment took 29 minutes per participant, on average. Prior to this experiment, a pilot test was performed with three participants to ensure that task descriptions are easy to understand and remember.

Measures and Analysis

For each user and visualization, we recorded eye tracking data, the task completion time, and whether the given response was correct. For each visualization, we additionally created a saliency map. From this raw data, we used the following measures in our analysis:

Correctness refers to the ratio of correctly answered questions for a given task per user. An answer was considered as correct when it contained all target labels or their values. Task correctness was checked manually after the experiment. We used the correctness to test if the complexity of the tasks was similar, and only included measures of correctly answered samples for further analysis.

From the recorded eye tracking data, we computed several fixation and AOI fixation measures:

To measure *fixation similarity* within and across tasks, we built a binary fixation map for each participant with ones at exact fixation locations and blurred the maps (Gaussian filter: size = 200, σ = 32). The *inter-participant fixation similarity* corresponds to the average value of *correlation coefficients (CC)* between each participant pair's fixation map solving the same task for the same visualization. This measure reveals the coherence of the fixations between users solving the same task (H1.1). The *inter-task fixation similarity* is the average of CC between each task pair's fixation map for the same visualization.

For the AOI fixation measures, we set the maximum distance between a fixation and an AOI to 50 px. This corresponds approximately to 1.3° of visual angle. The *first fixation time (FF)* – or time to first fixation – describes how much time passed from stimulus onset until the first fixation was registered within an AOI. The FF is used to compare the fixation sequence of task-dependent AOIs between tasks (H1.2).

To evaluate the prediction ability of saliency models and to measure the impact of saliency on attention (the *fixation-saliency similarity*), we generated saliency maps from 12 saliency algorithms, denoted **Itti** [96] (implementation by Harel [80]), **AIM** [24], **GBVS** [80], **SUN** [229], **CAS** [71], **Sign**[85], **BMS** [226], **eDN** [199], **SAMv** and **SAMr** [40] (feature maps extracted by the convolutional neural model based on VGG-16 [185] and ResNet-50 [82], respectively), **DVS** [152] (with the optimal weight of text saliency for MASSVIS database) and **TextS** [152] (text saliency of the DVS model separately).

To evaluate the models, we computed AUC and NSS scores. In addition, we report the score of human IO (see definitions of evaluation scores in Section 2.10). Fixation-similarity measures are compared between the three low-level analytical tasks and the Mem-task to verify hypothesis H1.3.

The *AOI saliency* is computed as the average saliency value in an AOI. We computed the correlation between the AOI saliency and its FF to test its visual search efficiency depending on its saliency value (H2.1). Also, we compared the AOI saliencies of target data points between the tasks to test if extreme data points are more salient (H2.2).

Finally, the *task completion time* (TCT) is measured after understanding of a task, from the initial display of a visualization to the press of the spacebar. We use TCT to test if the FE-task can be solved more efficiently (H2.3).

Eye-tracking data of our experiment are publicly available[24].

Apparatus

We recorded eye-tracking data using Tobii X2-60 eye-trackers at 60 Hz and processed by Tobii I-VT fixation filter.All stimuli were displayed on 24.1-inch monitors with a resolution

[24] `http://vgg.fiit.stuba.sk/2018-02/taskvis/`

of 1920 $\times$ 1080 pixels at a viewing distance of approximately 60 cm.

Participants

We recorded eye-tracking data, response times, and the textual responses of 47 students participating in a data visualization course and a computer vision course. Students were aged 20 to 25 years; 44 were male, three female. All participants gave their informed consent to the study and received an explanation of the experiment. The study was performed at the end of the course and participation was compulsory to gain all credits for the course. However, once having started the study, students were free to stop the experiment at any time without having their data recorded and losing any course credits.

3.7.4 Experimental Results

Each of the 47 users answered 30 questions in total. This corresponds to 10 answers per task for each participant, resulting in a total of 1410 gathered responses. From these responses, 199 (14.1%) were incorrectly answered and excluded from further analysis, leaving 1211 responses. While, in total, the highest number of incorrect answers was given for the F-task (81), there is no significant difference in *correctness* between the three tasks (Friedman test: $\chi^2(2) = 5.081, p = .079$)[12]. The difficulty of the three tasks therefore seems to be comparable.

Fixation Similarities

To test if fixation patterns are more coherent between users solving the same low-level task than when trying to memorize the visualized information, we compared the *inter-participant fixation similarity* between the three low-level analytical tasks, as well as the Mem-task. An ANOVA with Bonferroni-adjusted post-hoc comparisons showed that fixation similarity between participants is indeed significantly higher for the three analytical tasks of our experiment than for the Mem-task (as visualized in Figure 3.42; $F(3, 87) = 20.274; p < .001; \eta^2 = .411$). *This means that users solving the same low-level task indeed have more coherent fixations, thereby confirming hypothesis H1.1.*

To further explore this similarity between the fixations of the FE-task and the Mem-task, we compared the *inter-task fixation similarities* between all four tasks using an ANOVA with Bonferroni-corrected post-hoc comparisons. We found that the similarity between the FE-task and the Mem-task (rightmost bar in Figure 3.43) is significantly higher than between the F-task and the Mem-task, as well as the RV-task and the Mem-task ($F(5, 145) = 3.136; p = .010; \eta^2 = .098$). We did not find any statistical differences between inter-task fixation similarities of any other task pairs. In other words, the gaze patterns obtained during the Mem-task are indeed most similar to those of the FE-task, while the other two low-level analytical tasks lead to significantly less similar fixation patterns to the Mem-task.

Task-Dependent Fixation Sequence

We then tested if the high similarity between the fixation maps of users solving the same low-level analytical task can be explained by the sequence of fixations in the pre-defined

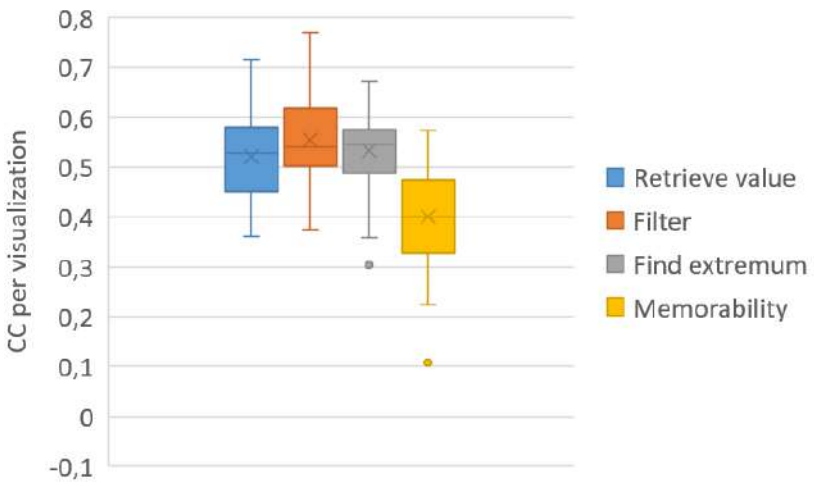


Figure 3.42: Similarity between fixations of the same type of activity.

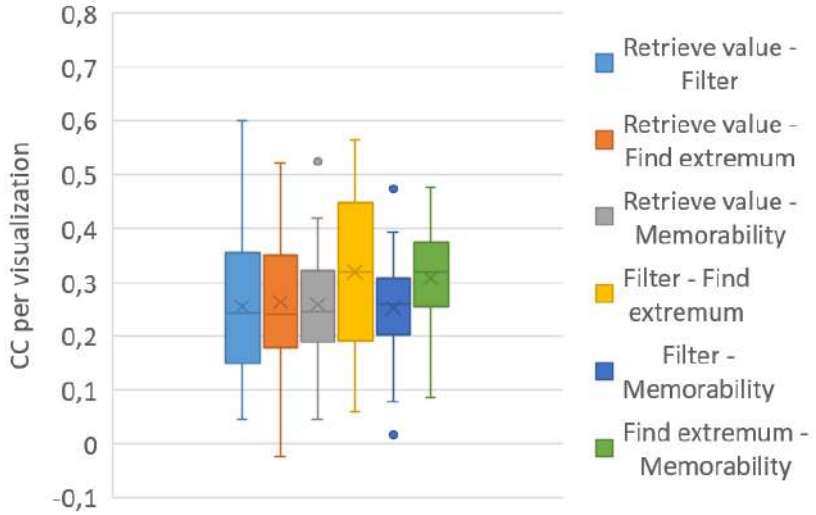


Figure 3.43: Similarity between fixations of different types of activity.

task-dependent AOIs. We therefore compared the first fixation times (FF) of the three task-dependent AOIs defined in Table 3.11 for each task type to test whether users follow the optimal viewing sequence (Figure 3.44). We conducted Friedman tests to compare the FFs on the target item label, data point, and value label, respectively, for the RV-task and F-task. For the FE-task, we conducted a Wilcoxon-Signed Rank test to compare the FFs on the target data point and its associated item label. We only found a significant difference in FFs for the RV-task ($\chi^2(2) = 111.972, p < .001$). Wilcoxon-Signed Rank pairwise post-hoc comparisons showed that all FFs were significantly different from each other, with the lowest FF for the item label, and the highest for the value label, as predicted in Table 3.11. The lowest median FF was recorded for data point in the FE-task, and interestingly also in the F-task. However, these differences are not statistically significant. *This only partially confirms hypothesis H1.2: while the task-dependent sequence of AOIs could predict the sequence of first fixations for the RV-task, this sequence could not be observed in the scanpaths recorded for the F-task, and is not pronounced for the FE-task.*

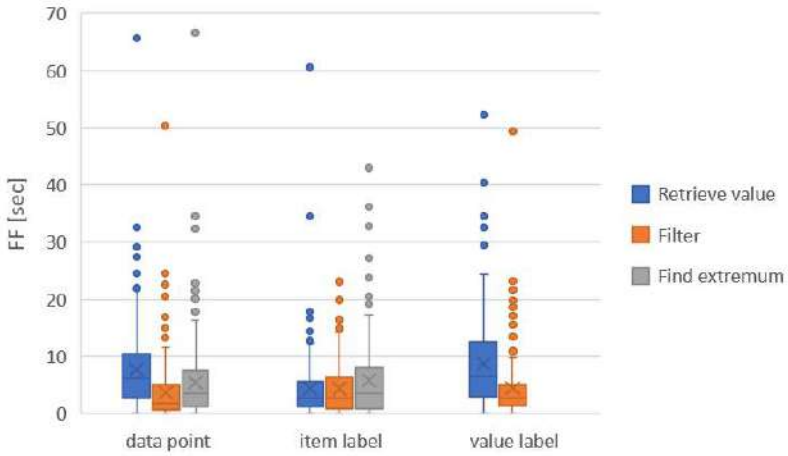


Figure 3.44: First fixation times of target-dependent AOIs defined in Table 3.11.

Fixation-Saliency Similarities

To test whether the *fixation-saliency similarity* is higher for the Mem-task than for the three low-level analytical tasks, we created saliency maps of all the 30 visualizations using 12 different algorithms and computed the fixation-saliency similarities for all four conditions.

In Table 3.12 and 3.13, we report average AUC and NSS scores of these saliency models and the *human IO*.

Comparing the performance of Itti et al.'s saliency model [96] to the human IO score in Tables 3.12 and 3.13, there is a remarkable gap between the saliency prediction and human visual attention for all four tasks. These scores are similar to the AUC and NSS scores of data visualization eye tracking data compared to Itti et al.'s [96] saliency model, reported by Haass et al. [79] (0.68 and 0.64, respectively). For reference, the average AUC and NSS scores they computed for natural images were 0.77 and 1.06, respectively.

We statistically compared the performance of two selected saliency models between the four tasks: the widely used saliency model by Itti et al. [96], as well as the state-of-the-art for modeling visual attention for visualizations [152] (DVS) using Kruskal-Wallis H tests. For both, AUC and NSS scores, we did not find any statistically significant differences between the tasks using Itti et al.'s saliency model (AUC: $\chi^2(3) = .017; p = .999$, NSS: $\chi^2(3) = .117; p = .990$). However, we found significant differences for DVS (AUC: $\chi^2(3) = 10.666; p = .014$, NSS: $\chi^2(3) = 16.972; p = .001$). Bonferroni-corrected Mann-Whitney U post-hoc comparisons showed a significantly higher AUC-score for the Mem-task than for the F-test and a significantly higher NSS-score for the Mem-task than all three low-level analytical tasks.

For the DVS model [152], we can therefore confirm our hypothesis H1.3 that bottom-up saliency strongly influences fixations of users when freely exploring the visualization, but has a significantly lower effect on visual attention when performing a low-level analytical task.

The major difference between the saliency model by Itti et al. [96] and DVS by Matzen et al. [152] is that the latter explicitly encodes text regions within visualizations as highly salient. A potential explanation for the significantly worse performance of DVS for the low-

Table 3.12: The average AUC scores for each task (saliency models sorted by publication year).

Task	**Itti**	**AIM**	**GBVS**	**SUN**	**CAS**	**Sign**	**BMS**	**eDN**	**SAMv**	**SAMr**	**TextS**	**DVS**	**IO**
RV	.684	.646	.608	.593	.595	.576	.621	.596	.630	.632	.647	**.702**	.812
F	.690	.645	.642	.593	.604	.622	.651	.595	.618	.631	.624	**.692**	.819
FE	.679	.654	.599	.602	.601	.600	.638	.568	.637	.647	.651	**.705**	.809
Mem	.686	.675	.553	.622	.637	.589	.652	.554	.653	.664	.696	**.738**	.781

Table 3.13: The average NSS scores for each task (saliency models sorted by publication year).

Task	**Itti**	**AIM**	**GBVS**	**SUN**	**CAS**	**Sign**	**BMS**	**eDN**	**SAMv**	**SAMr**	**TextS**	**DVS**	**IO**
RV	0.66	0.53	0.42	0.39	0.43	0.27	0.39	0.34	0.70	0.65	0.69	**0.80**	2.00
F	0.66	0.52	0.55	0.41	0.44	0.45	0.51	0.33	0.68	0.65	0.56	**0.71**	2.00
FE	0.64	0.55	0.41	0.45	0.46	0.35	0.48	0.24	0.80	0.78	0.72	**0.81**	1.93
Mem	0.67	0.62	0.24	0.53	0.63	0.34	0.54	0.19	0.83	0.88	1.03	**1.06**	1.50

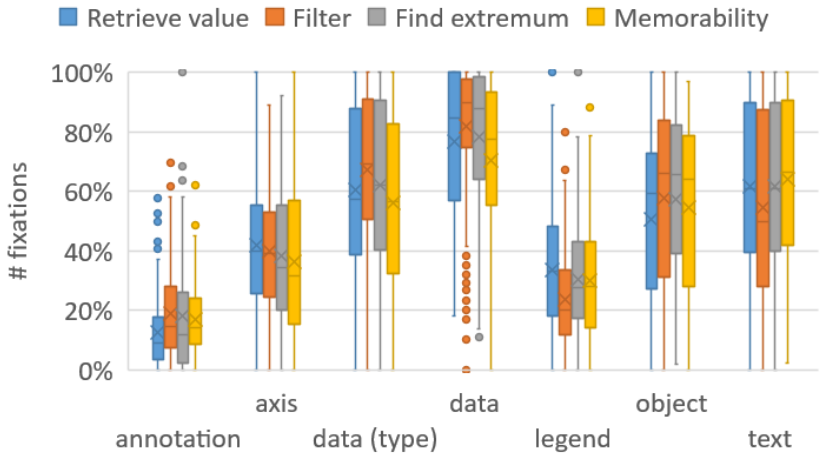


Figure 3.45: Fixations in task-independent AOIs.

level analytical tasks compared to the Mem-task could be that users direct their attention more towards the data areas than the text areas when performing low-level analytical tasks, than when trying to memorize the visualization. Therefore, we explored the *AOI fixation ratios* in task-independent AOIs defined in Table 3.9. Indeed, a Kruskal-Wallis H test with Mann-Whitney U post-hoc comparisons revealed that the data areas of visualizations were fixated more frequently during task-based analysis than during the memorability experiment ($\chi^2(3) = 41.435; p < .001$; see Figure 3.45). The reason for this could be that users were seeking more explicitly for a particular data point and spent less time reading annotations, legends, and titles to memorize textual information, which was irrelevant for the present task. For both, text elements and legends, the fixation ratio was significantly lower for the F-task than for all other tasks (Kruskal-Wallis H test: $\chi^2(3) = 23.701; p < .001$ and $\chi^2(3) = 37.121; p < .001$). However, there is no significant difference between the Mem-task and the other two low-level analytical tasks.

Correlation between Target Point Saliency and First Fixation Time

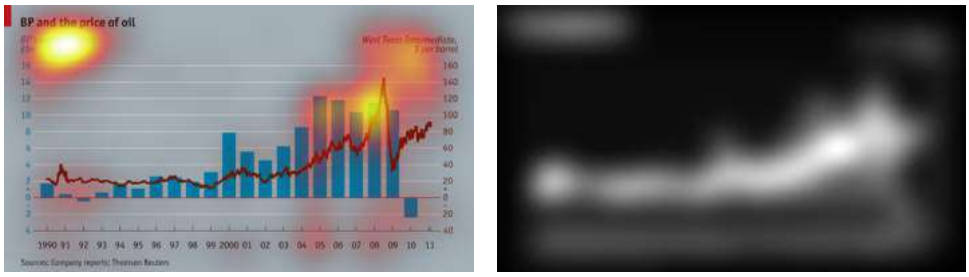
To test the influence of a target area's visual saliency on visual search performance, we computed the correlation between the task-dependent *AOI saliency* (using the model by Itti et al. [96]) and its *FF* for each of the three low-level analytical tasks. We only analyzed the first task-dependent AOI to be fixated according to the task-dependent AOI sequence shown as Step 1 in Table 3.11.

Since each task has a different optimal solution process, we set the target item label as visual search target for the RV-task, the target value label for the F-task, and the target data point for the FE-task. According to our hypothesis H2.1, there should be a negative correlation between the visual search target's AOI saliency and its FF – in other words: the more salient the target, the faster it should be fixated by the user. We found out that there is a negative correlation, but this correlation is weak (RV-task: $r = -0.26$, $p < .001$; F-task: $r = -0.21$, $p < .001$; FE-task: $r = -0.26$, $p < .001$). *In other words, the visual search efficiency for a target in the course of a low-level analytical task does not strongly correlate with its target saliency. Therefore, we have to reject hypothesis H2.1.*

Saliency of Extreme Data Points

Our assumption is that data points with extreme values usually stand out visually, i.e., have a higher saliency than target data points for the RV-task or the F-task. However, a Friedman test on the *AOI saliency* values computed using Itti et al.'s saliency model showed that there is, in fact, no difference in target data point saliency ($\chi^2(2) = 2.381; p = .304$). *Therefore, we have to reject hypothesis H2.2: target data points of the FE-task do not show up as more salient in saliency maps than target data points of the other two tasks.*

For illustration of this result, consider Figure 3.46: The peak in the line chart intuitively stands out. However, neither is this peak the most salient region in the saliency map, nor is it the extremum that was requested in the task question, which is the highest blue bar.



(a) When did BP achieve the highest net profit?

(b) Saliency map [96].

Figure 3.46: A fixation heatmap for the FE-task and a corresponding saliency map.

Find-Extremum Task Efficiency

Finally, we tested the hypothesis that extrema can be found more efficiently than values associated with given items or items associated with given value ranges. We therefore compared the task completion time between the three low-level analytical tasks. A Friedman test showed that there is a significant difference between the tasks: $\chi^2(2) = 16.128, p < .001$. Post-hoc Wilcoxon Signed-Rank tests revealed that the F-task takes significantly longer to be solved than the RV-task ($Z = -3.757, p < .001$) and the FE-task ($Z = -4.011, p < .001$), but there is no difference in task completion time between the RV-task and the FE-task ($Z = -.328, p = .743$). *We therefore also have to reject hypothesis 2.3: the FE-task was not more efficient to solve than the RV-task.*

3.7.5 Discussion

For our discussion, we will relate our results to our hypotheses.

Influence of Bottom-Up Saliency During Task-Based Visual Analysis

Fixation patterns of users solving different low-level analytical tasks showed that fixations between users solving the same task highly correlate, while the fixation map correlation between users trying to memorize the visualization is significantly lower. This result was

expected (**H1.1**). However only in the RV-task could we show that users clearly fixated the areas of interest in the optimal sequence for solving the task. The given low-level analytical task therefore seems to have a measurable top-down guidance for the users *where* to look, but not necessarily in which order (**H1.2**).

An unexpected finding during our experiment was that fixation maps of the memorability experiment much closer resembled those of the find-extremum task than those of the retrieve-value or filter task. We found that target data points in our FE-tasks did not have a higher bottom-up saliency than those of the RV-task and the F-task (**H2.2**).

An alternative explanation is that users were intentionally seeking for extrema as representative values to memorize the content of the visualization. This tendency is reflected in the selected descriptions of visualization content of users in Borkin et al.'s [20] memorability experiment. Most of the listed user descriptions contain a short summary of what is visualized together with one or more extreme items. This would mean that a memorability task would lead to similar top-down guidance as a find-extremum task.

Despite the higher diversity of the fixations during the memorability experiment, the fixations are more likely to co-incide with highly salient regions than during low-level analytical tasks. This is true for the DVS model recently presented by Matzen et al. [152] (**H1.3**). However, for the seminal saliency model by Itti et al. [96], the fixation-saliency similarities are equally low for the low-level analytical tasks as for the memorability task. The difference between these two saliency models is that DVS explicitly detects regions containing text and marks these regions as highly salient. Since it has been found that users attend textual elements for a long time when trying to memorize or free-viewing a visualization [153], DVS can better predict fixation patterns while performing exploratory visual analysis. During task-based visual analysis, however, users' attention is more strongly directed towards the data areas of the visualization, while, presumably, text areas are targeted only selectively.

Influence of Bottom-Up Saliency During Visual Search

Our second hypothesis was that bottom-up visual saliency may be a useful tool to predict the efficiency of visual search that needs to be conducted in the course of task-based visual analysis. Depending on the task, this may be a visual search for an item label, a value label, or an extreme data point. Especially in the latter case, we expected to see that extreme data points stand out in the saliency maps, and that therefore find-extremum tasks are more efficient to solve overall.

However, we could neither find a strong correlation between the target point's saliency and its first fixation time (H2.1), nor a significantly higher saliency of extreme data points compared to data point of the other the two tasks (H2.2), nor an increased task efficiency for the find-extremum task (H2.3). In other words: extreme data points are not necessarily more salient in classic visualizations, and more salient data points are not necessarily faster to detect during task-based visual analysis.

Study Limitations

Our study highlighted some unknown aspects about visual attention during task-based visual analysis. However, there are some limitations to our study.

First, the set of visualizations and the task questions were quite heterogeneous. The disadvantage is that there are, therefore, many factors potentially confounding the results, such as different visualization types, varying numbers of dependent variables encoded in the visualizations, or the usage of human recognizable objects. We suspect that some hypotheses would require a more controlled setup to be fully verified.

Second, our user group was composed of data visualization students who have gained some experience in analyzing visualization compared to novices. As also shown in a prior study [154], it can be expected that novices are less strongly guided by top-down factors when performing confirmatory analysis than more experienced users.

Third, we only tested three very simple analytical tasks in our experiment. However, in the task taxonomy by Amar et al. [6], there are more low-level tasks, like sort, determine range, cluster, or characterize distribution.

3.7.6 Summary

Our results show that despite having improved bottom-up saliency models for information visualization, like DVS [152], the influence of bottom-up visual saliency is drastically reduced during task-based visual analysis. We showed that users focus more on data areas of the visualization during task-based visual analysis than when trying to memorize a visualization. Therefore, the added text saliency in the DVS model did not increase the accuracy for task-based visual analysis in the same extent as for exploratory visual analysis. However, despite the increased attention in the data area, we did not find a strong correlation between a task-dependent area's saliency and visual search efficiency.

This means that visual attention is only slightly affected by early features when performing task-based visual analysis using information visualization – in contrast to observation of natural images or during exploratory visual analysis. Yet, fixations between users are more similar than during the memorability experiment. This means that task-based visual analysis is strongly guided by top-down factors imposed by the task.

To improve existing saliency models and tailor them more towards task-based visual analysis, we therefore recommended to merge classic image-based saliency models with object-based saliency models. When quantifying how much individual graphical marks stand out from their surrounding marks, the model should localize and identify the marks, compare their features at object level (e.g. color, orientation, size and shape) and estimate their relationships. In addition, other element types in a visualization, such as text areas, legends or axes, should be also incorporated in the model. For instance, saliency of text labels should vary with the task, so that text at informative locations for a given task receives higher saliency.

In future, it will be important to perform more systematic comparisons of eye tracking patterns between low-level analytical tasks. Since neural network models successfully reduced the gap between human fixations and saliency prediction for natural scenes, a similar approach could be applied in data visualizations, too. By training a neural network on viewers' fixation data acquired during a specific task, saliency maps could be tailored for a particular viewer and task respectively.

Chapter 4

Conclusions

This thesis explored various bottom-up and top-down factors of visual attention. We performed novel eye-tracking experiments, proposed computational saliency models and discussed human gaze behaviour. Visual attention modelling needs to have specialized fixation datasets which could improve saliency prediction. Therefore, we made our fixation databases available to the public.

Our experiments examined attentional factors separately. Most of them showed a high diversity of their effects on visual attention and visual performance, particularly in natural environments from the first-person perspective.

We explored salient 2D features such as color and shape using own synthetic image datasets. We found out that color contrasts in the LAB color space do not affect attention of individuals equally and therefore other high-level aspects of colors could be considered too. However, learned color associations to life-threatening situations, such as red indicating danger, have only a negligible effect on selective attention. Next, we showed that human gaze is directed to contour contrasts. This saliency is estimated with high precision using the spectral residual of the centroid distance signature. In addition, larger objects are fixated more frequently.

Furthermore, we pointed out that egocentric attention in real environments differs from image viewing conditions and therefore specialized saliency models need to be proposed. We observed that binocular vision is biased towards objects which are placed farther from an observer and neighboring objects, in contrast to images where the closest objects are highly fixated. While moving and surprising objects may significantly attract attention, we found that static factors dominate over dynamic ones. Our experiments revealed the highest saliency effects resulted from intensity, color and orientation contrasts. We also showed that egocentric vision is biased to the viewer's central visual field.

Beside low-level visual features, attention may be affected by a current emotional state of individuals. We observed that positive emotions produce a stronger saliency effect only during free viewing of valence-neutral stimuli. However, the opposite effect was observed during task-based analysis. We also found that tasks could be solved less efficiently when experiencing a positive mood and therefore, we suggest that it rather distracts users from a task.

Computational saliency models have been applied in many areas of computer science, mainly for natural images where neural network architectures significantly reduced the gap between predicted saliency and human fixations. These models have been also used in visualization

research, such as a quality metric. Because of notable difference between natural and synthetic images, our last research was focused on information visualizations and task-based visual analysis. The results showed that the effect of top-down attention significantly increases during various visual search tasks. We concluded that specialized computational models are needed for visualizations. In contrast to natural images, text saliency in visualizations has a great influence on attention.

In future work, the above mentioned experimental findings should be merged to create a computational model that could reliably predict attention in natural scenes and specialized domains, such as information visualizations or medical imaging. Because of the individuality in visual information processing, a possible solution could be to learn fixation preference from fixation data of a particular user solving a particular task using deep neural networks.

Bibliography

[1] R. Achanta et al. “SLIC superpixels compared to state-of-the-art superpixel methods”. In: *IEEE transactions on pattern analysis and machine intelligence* 34.11 (2012), pp. 2274–2282.

[2] R. Achanta and S. Süsstrunk. “Saliency detection for content-aware image resizing”. In: *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE. 2009, pp. 1005–1008.

[3] R. Achanta et al. “Frequency-tuned salient region detection”. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*. IEEE. 2009, pp. 1597–1604.

[4] R. Achanta et al. “Salient region detection and segmentation”. In: *International conference on computer vision systems*. Springer. 2008, pp. 66–75.

[5] R. Allison, B. Gillam, and E. Vecellio. “Binocular depth discrimination and estimation beyond interaction space”. In: *Journal of Vision* 7.9 (2007), pp. 817–817.

[6] R. Amar, J. Eagan, and J. Stasko. “Low-level components of analytic activity in information visualization”. In: *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*. IEEE. 2005, pp. 111–117.

[7] E. Ardizzone, A. Bruno, and G. Mazzola. “Saliency based image cropping”. In: *International Conference on Image Analysis and Processing*. Springer. 2013, pp. 773–782.

[8] A. Barbot and M. Carrasco. “Emotion and anxiety potentiate the way attention alters visual appearance”. In: *Scientific reports* 8.1 (2018), p. 5938.

[9] M. W. Becker and M. Leinenger. “Attentional selection is biased toward mood-congruent stimuli.” In: *Emotion* 11.5 (2011), p. 1248.

[10] M. Behrisch et al. “Quality Metrics for Information Visualization”. In: *Computer Graphics Forum*. Vol. 37. 3. Wiley Online Library. 2018, pp. 625–662.

[11] S. Belongie, J. Malik, and J. Puzicha. “Matching shapes”. In: *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. Vol. 1. IEEE. 2001, pp. 454–461.

[12] A. Betancourt et al. “An overview of first person vision and egocentric video analysis for personal mobile wearable devices”. In: *arXiv preprint arXiv* 1409 (2014).

[13] M. D. Binder, N. Hirokawa, and U. Windhorst. “Encyclopedia of neuroscience”. In: (2009).

[14] A. Borji and L. Itti. “Cat2000: A large scale fixation dataset for boosting saliency research”. In: *arXiv preprint arXiv:1505.03581* (2015).

[15] A. Borji and L. Itti. “Exploiting local and global patch rarities for saliency detection”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 478–485.

[16] A. Borji and L. Itti. “State-of-the-art in visual attention modeling”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2013), pp. 185–207.

[17] A. Borji, D. N. Sihite, and L. Itti. “Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study”. In: *IEEE Transactions on Image Processing* 22.1 (2013), pp. 55–69.

[18] A. Borji et al. “Analysis of scores, datasets, and models in visual saliency prediction”. In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 921–928.

[19] A. Borji et al. “Salient object detection: A survey. arXiv preprint”. In: *arXiv preprint arXiv:1411.5878* 2.4 (2014).

[20] M. A. Borkin et al. “Beyond memorability: Visualization recognition and recall”. In: *IEEE transactions on visualization and computer graphics* 22.1 (2016), pp. 519–528.

[21] M. A. Borkin et al. “What makes a visualization memorable?” In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2306–2315.

[22] M. Brehmer and T. Munzner. “A multi-level typology of abstract visualization tasks”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2376–2385.

[23] N. D. Bruce and J. K. Tsotsos. “Saliency, attention, and visual search: An information theoretic approach”. In: *Journal of vision* 9.3 (2009), pp. 5–5.

[24] N. Bruce and J. Tsotsos. “Saliency based on information maximization”. In: *Advances in neural information processing systems*. 2006, pp. 155–162.

[25] C. Bundesen and T. Habekost. *Principles of Visual Attention: Linking Mind and Brain*. Oxford University Press Oxford, 2008.

[26] M. Burch et al. “Evaluation of traditional, orthogonal, and radial tree diagrams by an eye tracking study”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2440–2448.

[27] M. Burch et al. “Visual task solution strategies in tree diagrams”. In: *2013 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE. 2013, pp. 169–176.

[28] N. J. Butko et al. “Visual saliency model for robot cameras”. In: *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE. 2008, pp. 2398–2403.

[29] Z. Bylinskii et al. “Eye fixation metrics for large scale evaluation and comparison of information visualizations”. In: *Workshop on Eye Tracking and Visualization*. Springer. 2015, pp. 235–255.

[30] Z. Bylinskii et al. “Learning visual importance for graphic designs and data visualizations”. In: *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. ACM. 2017, pp. 57–69.

[31] Z. Bylinskii et al. “What do different evaluation metrics tell us about saliency models?” In: *IEEE transactions on pattern analysis and machine intelligence* (2018).

[32] Z. Bylinskii et al. “Where should saliency models look next?” In: *European Conference on Computer Vision*. Springer. 2016, pp. 809–824.

[33] M. Cerf et al. “Predicting human gaze using low-level saliency combined with face detection”. In: *Advances in neural information processing systems*. 2008, pp. 241–248.

[34] C. Chamaret et al. “Adaptive 3D rendering based on region-of-interest”. In: *Stereoscopic Displays and Applications XXI*. Vol. 7524. International Society for Optics and Photonics. 2010, p. 75240V.

[35] A. Chambolle and T. Pock. “A first-order primal-dual algorithm for convex problems with applications to imaging”. In: *Journal of mathematical imaging and vision* 40.1 (2011), pp. 120–145.

[36] J. Chen and L. Zhang. “Joint multi-image saliency analysis for region of interest detection in optical multispectral remote sensing images”. In: *Remote Sensing* 8.6 (2016), p. 461.

[37] Y. Chen, C.-P. Yu, and G. Zelinsky. “Adding Shape to Saliency: A Proto-object Saliency Map for Predicting Fixations during Scene Viewing”. In: *Journal of Vision* 16.12 (2016), pp. 1309–1309.

[38] M.-M. Cheng et al. “Global contrast based salient region detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.3 (2015), pp. 569–582.

[39] R. J. Compton. “The interface between emotion and attention: A review of evidence from psychology and neuroscience”. In: *Behavioral and cognitive neuroscience reviews* 2.2 (2003), pp. 115–129.

[40] M. Cornia et al. “Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model”. In: *CoRR* abs/1611.09571 (2016).

[41] X. Cui, Q. Liu, and D. Metaxas. “Temporal spectral residual: fast motion saliency detection”. In: *Proceedings of the 17th ACM international conference on Multimedia*. ACM. 2009, pp. 617–620.

[42] A. M. Derrington, J. Krauskopf, and P. Lennie. “Chromatic mechanisms in lateral geniculate nucleus of macaque.” In: *The Journal of physiology* 357.1 (1984), pp. 241–265.

[43] K. Desingh et al. “Depth really Matters: Improving Visual Salient Region Detection with Depth.” In: *BMVC*. 2013.

[44] S. Dickinson and Z. Pizlo. *Shape perception in human and computer vision*. Springer, 2015.

[45] D. M. Diez, C. D. Barr, and M. Cetinkaya-Rundel. *OpenIntro statistics*. OpenIntro, 2012.

[46] X. Ding et al. “A Novel Emotional Saliency Map to Model Emotional Attention Mechanism”. In: *International Conference on Multimedia Modeling*. Springer. 2016, pp. 197–206.

[47] L. Dong et al. “Selective rendering with graphical saliency model”. In: *IVMSP Workshop, 2011 IEEE 10th*. IEEE. 2011, pp. 159–164.

[48] L. Duan et al. “Visual saliency detection by spatially weighted dissimilarity”. In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE. 2011, pp. 473–480.

[49] P. Ekman, W. V. Friesen, and P. Ellsworth. “Emotion in the human face: Guidelines for research and a review of findings”. In: *New York. Permagon* (1972).

[50] A. J. Elliot. “Color and psychological functioning: a review of theoretical and empirical work”. In: *Frontiers in Psychology* 6 (2015), p. 368.

[51] A. J. Elliot et al. “Color and psychological functioning: The effect of red on performance attainment.” In: *Journal of experimental psychology: General* 136.1 (2007), p. 154.

[52] D. Endres et al. “Hooligan detection: the effects of saliency and expert knowledge”. In: (2011).

[53] M. Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.

[54] S. Etchebehere and E. Fedorovskaya. “On the role of color in visual saliency”. In: *Electronic Imaging* 2017.14 (2017), pp. 58–63.

[55] M. Everingham et al. “The pascal visual object classes (voc) challenge”. In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[56] M. Everingham et al. “The pascal visual object classes challenge: A retrospective”. In: *International journal of computer vision* 111.1 (2015), pp. 98–136.

[57] M. W. Eysenck and M. T. Keane. *Cognitive psychology: A student’s handbook*. Psychology press, 2013.

[58] Y. Fang et al. “A video saliency detection model in compressed domain”. In: *IEEE transactions on circuits and systems for video technology* 24.1 (2014), pp. 27–38.

[59] A. Fathi, Y. Li, and J. M. Rehg. “Learning to recognize daily actions using gaze”. In: *European Conference on Computer Vision*. Springer. 2012, pp. 314–327.

[60] R. A. Ferrer et al. “The Effect of Emotion on Visual Attention to Information and Decision Making in the Context of Informed Consent Process for Clinical Trials”. In: *Journal of Behavioral Decision Making* 29.2-3 (2016), pp. 245–253.

[61] A. Field. *Discovering Statistics Using IBM SPSS Statistics: North American Edition*. Sage, 2017.

[62] B. L. Fredrickson. “The broaden-and-build theory of positive emotions.” In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 359.1449 (2004), p. 1367.

[63] B. L. Fredrickson and C. Branigan. “Positive emotions broaden the scope of attention and thought-action repertoires”. In: *Cognition & emotion* 19.3 (2005), pp. 313–332.

[64] D. Gao, V. Mahadevan, and N. Vasconcelos. “On the plausibility of the discriminant center-surround hypothesis for visual saliency”. In: *Journal of vision* 8.7 (2008), pp. 13–13.

[65] S. Garlandini and S. I. Fabrikant. “Evaluating the effectiveness and efficiency of visual variables for geographic information visualization”. In: *International Conference on Spatial Information Theory*. Springer. 2009, pp. 195–211.

[66] J. Gautier and O. Le Meur. “A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions”. In: *Cognitive Computation* 4.2 (2012), pp. 141–156.

[67] E. D. Gelasca, D. Tomasic, and T. Ebrahimi. “Which colors best catch your eyes: a subjective study of color saliency”. In: *First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*. Citeseer. 2005.

[68] J. J. Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.

[69] J. J. Gibson. *The perception of the visual world.* Houghton Mifflin, 1950.

[70] J. J. Gibson. *The senses considered as perceptual systems.* Houghton Mifflin, 1966.

[71] S. Goferman, L. Zelnik-Manor, and A. Tal. “Context-aware saliency detection”. In: *IEEE transactions on pattern analysis and machine intelligence* 34.10 (2012), pp. 1915–1926.

[72] J. H. Goldberg and J. I. Helfman. “Comparing Information Graphics: A Critical Look at Eye Tracking”. In: *Proceedings of the 3rd BELIV’10 Workshop: BEyond Time and Errors: Novel evaLuation Methods for Information Visualization*. BELIV ’10. Atlanta, Georgia: ACM, 2010, pp. 71–78. ISBN: 978-1-4503-0007-0.

[73] J. Goldberg and J. Helfman. “Eye tracking for visualization evaluation: Reading values on linear versus radial graphs”. In: *Information Visualization* 10.3 (2011), pp. 182–195.

[74] E. B. Goldstein. *Encyclopedia of perception*. Vol. 1. Sage, 2010.

[75] E. B. Goldstein and J. Brockmole. *Sensation and perception*. Cengage Learning, 2016.

[76] M. A. Goss-Sampson. *Statistical Analysis in JASP: A Guide for Students*. Version 2. 2018.

[77] R. L. Gregory. *Eye and brain: The psychology of seeing*. Princeton university press, 2015.

[78] M. Grol et al. “Effects of positive mood on attention broadening for self-related information”. In: *Psychological Research* 78.4 (2014), pp. 566–573.

[79] M. J. Haass et al. “Modeling human comprehension of data visualizations”. In: *International Conference on Virtual, Augmented and Mixed Reality*. Springer. 2016, pp. 125–134.

[80] J. Harel, C. Koch, and P. Perona. “Graph-based visual saliency”. In: *Advances in neural information processing systems*. 2007, pp. 545–552.

[81] S. Haroz and D. Whitney. “How capacity limits of attention influence information visualization effectiveness”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2402–2410.

[82] K. He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[83] H. v. Helmholtz. “On the theory of compound colors”. In: *Philosophical Magazine* 4 (1852), pp. 519–534.

[84] E. Hering. *Zur Lehre vom Lichtsinn*. Vol. 68. K. Akademie der Wissenschaften, 1878.

[85] X. Hou, J. Harel, and C. Koch. “Image signature: Highlighting sparse salient regions”. In: *IEEE transactions on pattern analysis and machine intelligence* 34.1 (2012), pp. 194–201.

[86] X. Hou and L. Zhang. "Saliency detection: A spectral residual approach". In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE. 2007, pp. 1–8.

[87] M.-K. Hu. "Visual pattern recognition by moment invariants". In: *IRE transactions on information theory* 8.2 (1962), pp. 179–187.

[88] Y. Hu, D. Rajan, and L.-T. Chia. "Adaptive local context suppression of multiple cues for salient visual attention detection". In: *2005 IEEE International Conference on Multimedia and Expo*. IEEE. 2005, p. 4.

[89] X. Huang et al. "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 262–270.

[90] L. M. Hurvich and D. Jameson. "An opponent-process theory of color vision." In: *Psychological review* 64.6p1 (1957), p. 384.

[91] A. W. Inhoff et al. "The size and direction of saccadic curvatures during reading". In: *Vision Research* 50.12 (2010), pp. 1117–1130. ISSN: 0042-6989.

[92] L. Itti and P. Baldi. "A principled approach to detecting surprising events in video". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE. 2005, pp. 631–637.

[93] L. Itti and A. Borji. "Computational models: Bottom-up and top-down aspects". In: *arXiv preprint arXiv:1510.07748* (2015).

[94] L. Itti, N. Dhavale, and F. Pighin. "Realistic avatar eye and head animation using a neurobiological model of visual attention". In: *Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI*. Vol. 5200. International Society for Optics and Photonics. 2003, pp. 64–79.

[95] L. Itti and C. Koch. "Computational modelling of visual attention". In: *Nature reviews neuroscience* 2.3 (2001), p. 194.

[96] L. Itti, C. Koch, and E. Niebur. "A model of saliency-based visual attention for rapid scene analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 20.11 (1998), pp. 1254–1259.

[97] H. Jänicke and M. Chen. "A Salience-based Quality Metric for Visualization". In: *Computer Graphics Forum*. Vol. 29. 3. Wiley Online Library. 2010, pp. 1183–1192.

[98] L. Jansen, S. Onat, and P. König. "Influence of disparity on fixation and saccades in free viewing of natural scenes". In: *Journal of Vision* 9.1 (2009), pp. 29–29.

[99] L. N. Jefferies et al. "Emotional valence and arousal interact in attentional control". In: *Psychological Science* 19.3 (2008), pp. 290–295.

[100] H. Jiang et al. "Automatic salient object segmentation based on context and shape prior." In: *BMVC*. Vol. 6. 7. 2011, p. 9.

[101] G. Johansson. "Visual motion perception". In: *Scientific American* 232.6 (1975), pp. 76–89.

[102] G. Johansson. "Visual perception of biological motion and a model for its analysis". In: *Perception & psychophysics* 14.2 (1973), pp. 201–211.

[103] R. Ju et al. “Depth saliency based on anisotropic center-surround difference”. In: *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE. 2014, pp. 1115–1119.

[104] T. Judd, F. Durand, and A. Torralba. “A benchmark of computational models of saliency to predict human fixations”. In: (2012).

[105] T. Judd et al. “Learning to predict where humans look”. In: *Computer Vision, 2009 IEEE 12th international conference on*. IEEE. 2009, pp. 2106–2113.

[106] B. Julesz. “Textons, the elements of texture perception, and their interactions”. In: *Nature* 290.5802 (1981), p. 91.

[107] A. Kaehler and G. Bradski. *Learning OpenCV 3: computer vision in C++ with the OpenCV library*. " O’Reilly Media, Inc.", 2016.

[108] J. W. Kalat. *Biological psychology*. Nelson Education, 2015.

[109] J. Karim, R. Weisz, and S. U. Rehman. “International positive and negative affect schedule short-form (I-PANAS-SF): Testing for factorial invariance across cultures”. In: *Procedia-Social and Behavioral Sciences* 15 (2011), pp. 2016–2022.

[110] K. Kaspar, R. R. Gameiro, and P. König. “Feeling good, searching the bad: Positive priming increases attention and memory for negative stimuli on webpages”. In: *Computers in Human Behavior* 53 (2015), pp. 332–343.

[111] W. D. Killgore and D. A. Yurgelun-Todd. “Positive affect modulates activity in the visual cortex to images of high calorie foods”. In: *International Journal of Neuroscience* 117.5 (2007), pp. 643–653.

[112] J. Kim and V. Pavlovic. “A shape preserving approach for salient object detection using convolutional neural networks”. In: *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE. 2016, pp. 609–614.

[113] S.-H. Kim et al. “Does an eye tracker tell the truth about visualizations?: findings while investigating visualizations for decision making”. In: *IEEE Transactions on Visualization & Computer Graphics* 12 (2012), pp. 2421–2430.

[114] Y. Kim and A. Varshney. “Saliency-guided enhancement for volume visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 12.5 (2006), pp. 925–932.

[115] L. A. King. *The science of psychology: An appreciative view*. McGraw-Hill Higher Education Boston, 2008.

[116] C. Koch and S. Ullman. “Shifts in selective visual attention: towards the underlying neural circuitry”. In: *Matters of intelligence*. Springer, 1987, pp. 115–141.

[117] K. Koffka. *Principles of Gestalt psychology*. Routledge, 2013.

[118] W. Köhler. “Gestalt psychology”. In: *Psychological research* 31.1 (1967), pp. XVIII–XXX.

[119] P. Kovesi. “Image segmentation using slic superpixels and dbscan clustering”. In: *University of Western Australia, Center for Exploration Targeting, Image Analysis Group* (2013).

[120] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[121] R. G. Kuehni and A. Schwarz. *Color ordered: a survey of color systems from antiquity to the present*. Oxford University Press, 2008.

[122] M. Kümmerer, L. Theis, and M. Bethge. "Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet". In: *arXiv preprint arXiv:1411.1045* (2014).

[123] M. Kümmerer et al. "Understanding low-and high-level contributions to fixation prediction". In: *2017 IEEE International Conference on Computer Vision*. 2017, pp. 4799–4808.

[124] K. Kurzhals et al. "Evaluating visual analytics with eye tracking". In: *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. ACM. 2014, pp. 61–69.

[125] C. Lang et al. "Depth matters: Influence of depth cues on visual saliency". In: *Computer vision–ECCV 2012*. Springer, 2012, pp. 101–115.

[126] G. Lara, A. De Antonio, and A. Peña. "A computational measure of saliency of the shape of 3D objects". In: *Trends and Applications in Software Engineering*. Springer, 2016, pp. 235–245.

[127] O. Le Meur and T. Baccino. "Methods for comparing scanpaths and saliency maps: strengths and weaknesses". In: *Behavior research methods* 45.1 (2013), pp. 251–266.

[128] S. Lee, M. Sips, and H.-P. Seidel. "Perceptually driven visibility optimization for categorical data visualization". In: *IEEE Transactions on visualization and computer graphics* 19.10 (2013), pp. 1746–1757.

[129] W. Lee et al. "Intelligent video surveillance system using dynamic saliency map and boosted Gaussian mixture model". In: *International Conference on Neural Information Processing*. Springer. 2011, pp. 557–564.

[130] J. Li and W. Gao. *Visual saliency computation: A machine learning perspective*. Vol. 8408. Springer, 2014.

[131] J. Li et al. "Estimating visual saliency through single image optimization". In: *IEEE Signal Processing Letters* 20.9 (2013), pp. 845–848.

[132] Y. Li, A. Fathi, and J. M. Rehg. "Learning to predict gaze in egocentric video". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 3216–3223.

[133] Y. Li et al. "The secrets of salient object segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 280–287.

[134] Y. Li et al. "Supervised saliency maps for first-person videos based on sparse coding". In: *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2018, pp. 2000–2005.

[135] T.-Y. Lin et al. "Microsoft COCO: Common objects in context". In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[136] D. T. Lindsey et al. "Color channels, not color appearance or color categories, guide visual search for desaturated color targets". In: *Psychological science* 21.9 (2010), pp. 1208–1214.

[137] H. Liu et al. "Improving visual saliency computing with emotion intensity". In: *IEEE transactions on neural networks and learning systems* 27.6 (2016), pp. 1201–1213.

[138] T. Liu et al. “Learning to detect a salient object”. In: *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. IEEE. 2007, pp. 1–8.

[139] Z. Liu, L. Meur, and S. Luo. “Superpixel-based saliency detection”. In: *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*. IEEE. 2013, pp. 1–4.

[140] Z. Liu et al. “Superpixel-based spatiotemporal saliency detection”. In: *IEEE transactions on circuits and systems for video technology* 24.9 (2014), pp. 1522–1540.

[141] Z. Liu et al. “Regions of interest extraction based on HSV color space”. In: *Industrial Informatics (INDIN), 2012 10th IEEE International Conference on*. IEEE. 2012, pp. 481–485.

[142] G. Loffler. “Perception of contours and shapes: Low and intermediate stage mechanisms”. In: *Vision research* 48.20 (2008), pp. 2106–2127.

[143] D. G. Lowe. “Object recognition from local scale-invariant features”. In: *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Vol. 2. Ieee. 1999, pp. 1150–1157.

[144] C. C. Loy, T. Xiang, and S. Gong. “Salient motion detection in crowded scenes”. In: *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on*. IEEE. 2012, pp. 1–4.

[145] T. Maekawa et al. “The effect of mood state on visual search times for detecting a target in noise: An application of smartphone technology”. In: *PloS one* 13.4 (2018), e0195865.

[146] V. Mahadevan and N. Vasconcelos. “Spatiotemporal saliency in dynamic scenes”. In: *IEEE transactions on pattern analysis and machine intelligence* 32.1 (2010), pp. 171–177.

[147] A. Maki, P. Nordlund, and J.-O. Eklundh. “A computational model of depth-based attention”. In: *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*. Vol. 4. IEEE. 1996, pp. 734–739.

[148] M. Mancas et al. *Computational attention towards attentive computers*. Presses univ. de Louvain, 2007.

[149] S. Marat et al. “Modelling spatio-temporal saliency to predict gaze direction for short videos”. In: *International journal of computer vision* 82.3 (2009), p. 231.

[150] L. Marchesotti, C. Cifarelli, and G. Csurka. “A framework for visual saliency detection with applications to image thumbnailing”. In: *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE. 2009, pp. 2232–2239.

[151] K. Matsuo et al. “An attention-based activity recognition for egocentric video”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014, pp. 551–556.

[152] L. E. Matzen et al. “Data Visualization Saliency Model: A Tool for Evaluating Abstract Data Visualizations”. In: *IEEE Transactions on Visualization & Computer Graphics* 1 (2018), pp. 563–573.

[153] L. E. Matzen et al. “Patterns of attention: How data visualizations are read”. In: *International Conference on Augmented Cognition*. Springer. 2017, pp. 176–191.

[154] L. E. Matzen et al. “Using eye tracking metrics and visual saliency maps to assess image utility”. In: *Electronic Imaging* 2016.16 (2016), pp. 1–8.

[155] R. Mazza. *Introduction to information visualization*. Springer Science & Business Media, 2009.

[156] T. Mei et al. “VideoSense: towards effective online video advertising”. In: *Proceedings of the 15th ACM international conference on Multimedia*. ACM. 2007, pp. 1075–1084.

[157] S. Miyazawa and S. Iwasaki. “Effect of negative emotion on visual attention: Automatic capture by fear-related stimuli”. In: *Japanese Psychological Research* 51.1 (2009), pp. 13–23.

[158] S. B. Most et al. “The naked truth: Positive, arousing distractors impair rapid target perception”. In: *Cognition and emotion* 21.5 (2007), pp. 964–981.

[159] M. C. Mozer and S. P. Vecera. “Space- and object-based attention”. In: *Neurobiology of attention*. Elsevier, 2005, pp. 130–134.

[160] Y. Niu et al. “Leveraging stereopsis for saliency analysis”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 454–461.

[161] A. Öhman, A. Flykt, and F. Esteves. “Emotion drives attention: detecting the snake in the grass.” In: *Journal of experimental psychology: general* 130.3 (2001), p. 466.

[162] C. N. Olivers and S. Nieuwenhuis. “The beneficial effect of concurrent task-irrelevant mental activity on temporal attention”. In: *Psychological science* 16.4 (2005), pp. 265–269.

[163] N. Ouerhani and H. Hugli. “Computing visual attention from scene depth”. In: *Pattern Recognition, 2000. Proceedings. 15th International Conference on*. Vol. 1. IEEE. 2000, pp. 375–378.

[164] S. Palmisano et al. “Stereoscopic perception of real depths at large distances”. In: *Journal of vision* 10.6 (2010), pp. 19–19.

[165] D. Parkhurst, K. Law, and E. Niebur. “Modeling the role of salience in the allocation of overt visual attention”. In: *Vision research* 42.1 (2002), pp. 107–123.

[166] C. Pêcher, C. Lemercier, and J.-M. Cellier. “Emotions drive attention: Effects on driver’s behaviour”. In: *Safety Science* 47.9 (2009), pp. 1254–1259.

[167] K.-C. Peng et al. “Where do emotions come from? predicting the emotion stimuli map”. In: *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE. 2016, pp. 614–618.

[168] F. Perazzi et al. “Saliency filters: Contrast based filtering for salient region detection”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 733–740.

[169] R. J. Peters et al. “Components of bottom-up gaze allocation in natural images”. In: *Vision Research* 45.18 (2005), pp. 2397–2416. ISSN: 0042-6989.

[170] E. A. Phelps, S. Ling, and M. Carrasco. “Emotion facilitates perception and potentiates the perceptual benefits of attention”. In: *Psychological science* 17.4 (2006), pp. 292–299.

[171] M. Pohl, M. Schmitt, and S. Diehl. "Comparing the Readability of Graph Layouts Using Eyetracking and Task-oriented Analysis". In: *Proceedings of the Fifth Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*. Computational Aesthetics'09. Victoria, British Columbia, Canada: Eurographics Association, 2009, pp. 49–56. ISBN: 978-3-905674-17-0.

[172] J. Qi et al. "Saliency detection via joint modeling global shape and local consistency". In: *Neurocomputing* 222 (2017), pp. 81–90.

[173] C. Ramasamy et al. "Using eye tracking to analyze stereoscopic filmmaking". In: *SIGGRAPH'09: Posters*. ACM. 2009, p. 28.

[174] K. Rapantzikos et al. "Bottom-up spatiotemporal visual attention model for video analysis". In: *IET Image Processing* 1.2 (2007), pp. 237–248.

[175] M. Raschke et al. "Visual analysis of perceptual and cognitive processes". In: *Information Visualization Theory and Applications (IVAPP), 2014 International Conference on*. IEEE. 2014, pp. 284–291.

[176] J. E. Raymond, M. J. Fenske, and N. T. Tavassoli. "Selective attention determines emotional responses to novel visual stimuli". In: *Psychological science* 14.6 (2003), pp. 537–542.

[177] J. E. Raymond, K. L. Shapiro, and K. M. Arnell. "Temporary suppression of visual processing in an RSVP task: An attentional blink?" In: *Journal of experimental psychology: Human perception and performance* 18.3 (1992), p. 849.

[178] A. Recasens et al. "Towards cognitive saliency: narrowing the gap to human performance". In: *Journal of Vision* 17.10 (2017), pp. 542–542.

[179] N. Riche et al. "Saliency and human fixations: state-of-the-art and study of comparison metrics". In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 1153–1160.

[180] G. Rowe, J. B. Hirsh, and A. K. Anderson. "Positive affect increases the breadth of attentional selection". In: *Proceedings of the National Academy of Sciences* 104.1 (2007), pp. 383–388.

[181] A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner. "Eye movements and perception: A selective review". In: *Journal of vision* 11.5 (2011), p. 9.

[182] C. Shen and Q. Zhao. "Webpage saliency". In: *European conference on computer vision*. Springer. 2014, pp. 33–46.

[183] C. Siagian and L. Itti. "Biologically inspired mobile robot vision localization". In: *IEEE Transactions on Robotics* 25.4 (2009), pp. 861–873.

[184] H. Siirtola et al. "Visual perception of parallel coordinate visualizations". In: *Information Visualisation, 2009 13th International Conference*. IEEE. 2009, pp. 3–9.

[185] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[186] D. Smilek et al. "Relax! Cognitive strategy influences visual search". In: *Visual Cognition* 14.4-8 (2006), pp. 543–564.

[187] R. Song et al. "Conditional random field-based mesh saliency." In: *ICIP*. 2012, pp. 637–640.

[188] M. Sonka, V. Hlavac, and R. Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[189] N. Sprague and D. Ballard. "Eye movements for reward maximization". In: *Advances in neural information processing systems*. 2004, pp. 1467–1474.

[190] M. Sun et al. "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction". In: *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2016, pp. 1–6.

[191] C. Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[192] J. M. Talarico, D. Berntsen, and D. C. Rubin. "Positive emotions enhance recall of peripheral details". In: *Cognition and Emotion* 23.2 (2009), pp. 380–398.

[193] H. R. Tavakoli et al. "Digging Deeper Into Egocentric Gaze Prediction". In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 273–282.

[194] A. Torralba et al. "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search." In: *Psychological review* 113.4 (2006), p. 766.

[195] A. Treisman. "Feature binding, attention and object perception". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 353.1373 (1998), pp. 1295–1306.

[196] A. M. Treisman and G. Gelade. "A feature-integration theory of attention". In: *Cognitive psychology* 12.1 (1980), pp. 97–136.

[197] A. Treisman and S. Gormican. "Feature analysis in early vision: evidence from search asymmetries." In: *Psychological review* 95.1 (1988), p. 15.

[198] R. Valenti, N. Sebe, and T. Gevers. "Image saliency by isocentric curvedness and color". In: *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE. 2009, pp. 2185–2192.

[199] E. Vig, M. Dorr, and D. Cox. "Large-scale optimization of hierarchical features for saliency prediction in natural images". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 2798–2805.

[200] H. A. Wadlinger and D. M. Isaacowitz. "Positive mood broadens visual attention to positive stimuli". In: *Motivation and Emotion* 30.1 (2006), pp. 87–99.

[201] J. Wang et al. "Computational model of stereoscopic 3D visual saliency". In: *IEEE Transactions on Image Processing* 22.6 (2013), pp. 2151–2165.

[202] J. Wang et al. "Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli". In: *Journal of Eye Movement Research* 5.5 (2012).

[203] M. O. Ward, G. Grinstein, and D. Keim. *Interactive data visualization: foundations, techniques, and applications*. AK Peters/CRC Press, 2015.

[204] E. R. Watkins and N. J. Moberly. "Concreteness training reduces dysphoria: A pilot proof-of-principle study". In: *Behaviour Research and Therapy* 47.1 (2009), pp. 48–53.

[205] D. Watson, L. A. Clark, and A. Tellegen. “Development and validation of brief measures of positive and negative affect: the PANAS scales.” In: *Journal of personality and social psychology* 54.6 (1988), p. 1063.

[206] T. P. Weldon, W. E. Higgins, and D. F. Dunn. “Gabor filter design for multiple texture segmentation”. In: *Optical Engineering* 35.10 (1996), pp. 2852–2864.

[207] M. Wertheimer. “Untersuchungen zur Lehre von der Gestalt”. In: *Psychologische Forschung* 1.1 (1922), pp. 47–58.

[208] J. M. Wolfe. “Guided search 2.0 a revised model of visual search”. In: *Psychonomic bulletin & review* 1.2 (1994), pp. 202–238.

[209] J. M. Wolfe. “Guided Search 4.0: A guided search model that does not require memory for rejected distractors”. In: *Journal of Vision* 1.3 (2001), pp. 349–349.

[210] J. M. Wolfe, K. R. Cave, and S. L. Franzel. “Guided search: an alternative to the feature integration model for visual search.” In: *Journal of Experimental Psychology: Human perception and performance* 15.3 (1989), p. 419.

[211] J. M. Wolfe and G. Gancarz. “Guided Search 3.0”. In: *Basic and clinical applications of vision science*. Springer, 1997, pp. 189–192.

[212] L. E. Wool et al. “Salience of unique hues and implications for color theory”. In: *Journal of vision* 15.2 (2015), pp. 10–10.

[213] J. Xu et al. “Gaze-enabled egocentric video summarization via constrained submodular maximization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 2235–2244.

[214] K. Yamada et al. “Attention prediction in egocentric video using motion and visual saliency”. In: *Pacific-Rim Symposium on Image and Video Technology*. Springer. 2011, pp. 277–288.

[215] K. Yamada et al. “Can saliency map models predict human egocentric visual attention?” In: *Asian Conference on Computer Vision*. Springer. 2010, pp. 420–429.

[216] M. Yang, K. Kpalma, and J. Ronsin. *A survey of shape feature extraction techniques*. 2008.

[217] A. L. Yarbus. “Eye movements during perception of complex objects”. In: *Eye movements and vision*. Springer, 1967, pp. 171–211.

[218] T. Young. “II. The Bakerian Lecture. On the theory of light and colours”. In: *Philosophical transactions of the Royal Society of London* 92 (1802), pp. 12–48.

[219] C. Zach, T. Pock, and H. Bischof. “A duality based approach for realtime TV-L 1 optical flow”. In: *Joint Pattern Recognition Symposium*. Springer. 2007, pp. 214–223.

[220] J. R. Zadra and G. L. Clore. “Emotion and perception: The role of affective information”. In: *Wiley interdisciplinary reviews: cognitive science* 2.6 (2011), pp. 676–685.

[221] Y. Zhai and M. Shah. “Visual attention detection in video sequences using spatiotemporal cues”. In: *Proceedings of the 14th ACM international conference on Multimedia*. ACM. 2006, pp. 815–824.

[222] D. Zhang and G. Lu. “Review of shape representation and description techniques”. In: *Pattern recognition* 37.1 (2004), pp. 1–19.

[223] D. Zhang, G. Lu, et al. “A comparative study on shape retrieval using Fourier descriptors with different shape signatures”. In: *Proc. of international conference on intelligent multimedia and distance education (ICIMADE01)*. 2001, pp. 1–9.

[224] H. Zhang et al. “Depth combined saliency detection based on region contrast model”. In: *Computer Science & Education (ICCSE), 2012 7th International Conference on*. IEEE. 2012, pp. 763–766.

[225] J. Zhang and S. Sclaroff. “Exploiting surroundedness for saliency detection: a boolean map approach”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 5 (2016), pp. 889–902.

[226] J. Zhang and S. Sclaroff. “Saliency detection: A boolean map approach”. In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 153–160.

[227] L. Zhang, L. Yang, and T. Luo. “Unified saliency detection model using color and texture features”. In: *PloS one* 11.2 (2016), e0149328.

[228] L. Zhang and W. Lin. *Selective visual attention: computational models and applications*. John Wiley & Sons, 2013.

[229] L. Zhang et al. “SUN: A Bayesian framework for saliency using natural statistics”. In: *Journal of vision* 8.7 (2008), pp. 32–32.

[230] M. Zhang et al. “Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4372–4381.

[231] Y. Zhang et al. “Stereoscopic visual attention model for 3D video”. In: *International Conference on Multimedia Modeling*. Springer. 2010, pp. 314–324.

[232] P. G. Zimbardo, R. Johnson, and V. McCann. *Psychology: Core Concepts*. Boston, MA: Pearson Education, Inc, 2012.

Relevant Publications of the Author

[233] V. Olesova, W. Benesova, and P. Polatsek. “Visual attention in egocentric field-of-view using RGB-D data”. In: *Ninth International Conference on Machine Vision (ICMV 2016)*. Vol. 10341. International Society for Optics and Photonics. 2017, 103410T.

[234] P. Polatsek and W. Benesova. “Bottom-up saliency model generation using superpixels”. In: *Proceedings of the 31st Spring Conference on Computer Graphics*. ACM. 2015, pp. 121–129.

[235] P. Polatsek et al. “Computational models of shape saliency”. In: *Eleventh International Conference on Machine Vision (ICMV 2018)*. Vol. 11041. International Society for Optics and Photonics. 2019, 110412B.

[236] P. Polatsek et al. “Exploring visual attention and saliency modeling for task-based visual analysis”. In: *Computers & Graphics* 72 (2018), pp. 26–38.

[237] P. Polatsek et al. “Novelty-based spatiotemporal saliency detection for prediction of gaze in egocentric video”. In: *IEEE Signal Processing Letters* 23.3 (2016), pp. 394–398.

Appendix A

DVD Contents

- `/motion/` is related to our contribution described Section 3.1
- `/color/` is related to our contribution described Section 3.2
- `/shape/` is related to our contribution described Section 3.3
- `/depth/` is related to our contribution described Section 3.4
- `/egocentric/` is related to our contribution described Section 3.5
- `/emotions/` is related to our contribution described Section 3.6
- `/infovis/` is related to our contribution described Section 3.7
- `/publications/` contains author's publications

Appendix B

Resumé

Úvod

Vizuálnu pozornosť tvorí séria kognitívnych procesov, ktoré vyberajú relevantné informácie z prostredia a potláčajú tie nepodstatné [13]. Preto pozornosť zohráva dôležitú úlohu pri koordinácii pohybov hlavy a očí. Pri skenovaní scény vykonávame sekvenciu rýchlych pohybov (sakád) a fixácií, kedy je oko relatívne v pokoji a prijíma vizuálnu informáciu [181, 75].

Pozornosť je ovplyvňovaná zdola nahor prostredníctvom nápadných stimulov ako aj zhora nadol našimi cieľmi a vedomosťami. Predpokladá sa, že tieto dva procesy spolu kooperujú a tým umožňujú vnímať prostredie [25].

Vo všeobecnosti existujú dva prístupy ako definovať nápadnosť. Môžeme merať nápadnosť napr. prostredníctvom snímania pohybov očí alebo môžeme predpovedať nápadnosť výpočtovými modelmi, ktoré vytvárajú mapy nápadnosti [148].

Výzvy

Hlavným cieľom tejto práce je výskum rôznych aspektov vizuálnej pozornosti prostredníctvom nových experimentov a výpočtových modelov nápadnosti.

Medzi hlavné výzvy tejto práce patrí:

1. **Egocentrická vizuálna pozornosť:** Pozornosť v reálnom prostredí sa odlišuje od pozornosti pri sledovaní klasických videí. Modelovanie nápadnosti vo videách z pohľadu prvej osoby by sa na rozdiel od štandardných modelov malo zamerať aj na pohyb pozorovateľa, binokulárnu hĺbku a predchádzajúce znalosti pozorovateľa o prostredí, v ktorom sa nachádza [75].
2. **Vizuálna pozornosť ovplyvnená 2D statickými príznakmi:** Modely založené na prístupe zdola nahor väčšinou používajú rozdiely v intenzite, farbe a orientácii pri predpovedaní nápadných oblastí [25, 75, 16]. Napriek tomu je dôležité pokračovať aj vo výskume farebného vnímania, aby sme pochopili ako pozornosť ovplyvňujú farebné rozdiely a asociácie s jednotlivými farbami (napr. červená farba spojená s nebezpečenstvom) [51]. Okrem toho pozornosť ovplyvňuje aj veľkosť a tvar objektov [44].

Modelovanie takejto nápadnosti by malo byť riešené lokálne na úrovni kontúr ako aj globálne na úrovni jednotlivých objektov [142, 13].

3. **Pozornosť pri riešení úloh vo vizualizáciách:** Väčšina modelov nápadnosti bola vyhodnocovaná na prirodzených scénach počas ich voľného prezerania. Modelovanie pozornosti sa už využíva tiež ako ukazovateľ kvality vizualizácií [10]. Medzi fotografiami a vizualizáciami je ale niekoľko rozdielov, ktoré je potrebné preskúmať.
4. **Vplyv nálady na vizuálnu pozornosť a hľadanie:** Teória rozšírenia hovorí, že pozitívne emócie rozširujú vizuálnu pozornosť [62]. Aj keď existujú modely nápadnosti, ktoré sú určené pre emocionálne stimuly, modely nápadnosti neboli doteraz nikdy vyhodnotené pri rôznych emóciách pozorovateľov.

Hlavné prínosy práce

Medzi hlavné prínosy tejto práce patrí:

1. **Výskum egocentrickej nápadnosti v reálnom prostredí** [234, 237, 233]:
   - analýza egocentrickej nápadnosti pohybu a prekvapenia prostredníctvom egocentrického datasetu,
   - analýza egocentrickej nápadnosti hĺbky vo vlastnom experimente a nových modelov nápadnosti,
   - analýza statických príznakov vo vlastnom egocentrickom experimente a nových modelov nápadnosti.
2. **Výskum vplyvu farby, tvaru a emócií pri prezeraní obrazov a možností ich zakomponovania v modeloch nápadnosti** [235]:
   - analýza nápadnosti tvaru vo vlastnom experimente a nových modelov nápadnosti,
   - analýza nápadnosti farby vo vlastnom experimente,
   - analýza pozornosti a vyhodnotenie modelu nápadnosti pod vplyvom emócií.
3. **Výskum pozornosti počas riešenia úloh vo vizualizáciách** [236]:
   - analýza riadenia pozornosti vytýčenými cieľmi vo vlastnom experimente a vyhodnotenie modelov nápadnosti vo vizualizáciách.

Analýza existujúcich prác

Predchádzajúce desaťročia výskumu vizuálnej pozornosti priniesli mnoho výpočtových modelov, ktoré môžme deliť podľa [16, 130]:

- **faktorov, ktoré priťahujú pozornosť:** faktory riadené stimulmi (*zdola nahor*), faktory riadené cieľmi (*zhora nadol*) a ich kombinácia,
- **časovej zložky:** *priestorové* modely založené výhradne na aktuálnej scéne a *časové* modely založené na predchádzajúcich vedomostiach, analýze pohybu alebo kombinácii oboch prístupov,

- **typu stimulov:** *statické* stimuly ako napr. intenzita, farba a hĺbka a *dynamické* stimuly ako pohyb a blikanie,
- **typu úloh:** *voľné prezeranie*, *vizuálne hľadanie* a iné *komplexnejšie úlohy*, napr. šoférovanie,
- **jednotky nápadnosti:** modely založené na *polohe*, ktoré priraďujú nápadnosť každej pozícii alebo modely založené na *objektoch*, ktoré získajú nápadné objekty z predchádzajúceho modelu alebo počítajú nápadnosť priamo na úrovni objektov,
- **množstva informácií** použitých pri predpovedaní nápadnosti: *lokálna* informácia len z časti obrazu alebo *globálna* informácia z celého obrazu.

Prínos práce

V tejto práci skúmame stimuly a ciele, ktoré ovplyvňujú pozornosť prostredníctvom nových modelov nápadnosti a experimentov so statickými obrazmi a egocentrickými videami. Výsledky týchto štúdií môžu prispieť k pochopeniu ľudskej vizuálnej pozornosti a k zlepšeniu predpovedania pozornosti.

Modelovanie egocentrickej nápadnosti pohybu

Keďže egocentrická nápadnosť bola doteraz skúmaná len okrajovo, v diplomovej práci sme navrhli časopriestorový model nápadnosti pre egocentrické videá, ktorý v tejto časti overíme na rozsiahlejšom datasete. Tento model bol publikovaný v [234] a [237].

Náš model nápadnosti [234, 237] využíva superpixelovú segmentáciu [1], čím aspoň čiastočne pokrýva pozornosť založenú na objektoch. Superpixel reprezentujeme statickými (intenzita, farba a orientácia) a dynamickými príznakmi (pohyb) na rôznych úrovniach detailu. Keďže ľudský pohľad smeruje tiež k neočakávaným a prekvapujúcim stimulom, náš model využíva aj prekvapenie v pohybe. Pre každý superpixel potom vytvárame histogramy týchto príznakov.

Podobne ako Itti a kol. [96] aj my vytvárame Gaussovu pyramídu, ale zo superpixelov. Následne porovnáme ich histogramy medzi jemnejšími a hrubšími vrstvami pyramídy, čím vypočítame priestorovú a časovú nápadnosť. V prípade pohybového prekvapenia porovnávame predchádzajúce znalosti o pohybe na danej pozícii s aktuálnou videosnímkou.

Dataset, ktorý sme použili na vyhodnotenie, pozostáva z 2 osôb, ktorých pohľad bol zaznamenaný v obchode, v ktorom mali za úlohu nájsť určité tovary. Okrem nášho modelu sme vyhodnotili aj existujúce modely – štandardný priestorový model [96] a časopriestorový superpixelový model [139].

Z výsledkov sme zistili, že statická nápadnosť dominuje nad dynamickou napriek tomu, že sa vo videách hýbu viaceré objekty. Výsledky tiež naznačujú, že tieto nápadnosti neovplyvňujú participantov rovnako a ich pozornosť mohla byť riadená aj hĺbkou a inými vysokoúrovňovými faktormi (napr. detekcia biologického pohybu).

Vizuálna pozornosť ovplyvnená farbou

Farba je základným prvkom vizuálnej pozornosti. Niektoré práce naznačujú, že okrem nápadnosti z farebného kontrastu by sa mala brať do úvahy aj psychológia farieb. Datasety špecializujúce sa na farbu nie sú verejne dostupné. Preto sme sa rozhodli náš nový experiment s farbou verejne sprístupniť. Pomocou neho sme zisťovali, či farebné rozdiely môžu spoľahlivo modelovať farebnú nápadnosť, alebo existujú farby, ktoré sú viac fixované než ostatné (napr. červená súvisiaca s nebezpečenstvom a žltá s upozornením).

V našom experimente sme zobrazovali rôznofarebné objekty na jednotnom pozadí 15 študentom, pričom sme využili farby ako červená, zelená, modrá, žltá, tyrkysová, purpurová, ružová a oranžová.

V získaných dátach sme našli vysokú koreláciu fixácii a farebného kontrastu pre použité obrazy. Avšak výsledky po jednotlivých participantoch ukázali, že pozornosť niektorých ľudí ovplyvňuje kontrast iba minimálne. Ďalej sme zistili, že farby, ktoré môžu byť spájané s nebezpečenstvom (červená a žltá) majú len miernu prevahu vo fixovaní. Na druhej strane sme našli výrazne menej fixácii tyrkysovej farby v porovnaní s ostatnými objektami.

Vizuálna pozornosť ovplyvnená tvarom

Okrem jednoduchých stimulov ako intenzita, farba a orientácia ovplyvňuje pozornosť aj veľkosť a tvar objektov. Preto sme sa rozhodli zistiť, či a v akom rozsahu ovplyvňujú pozornosť lokálne a globálne charakteristiky tvaru objektov a ich vzájomné rozdiely. Keďže dataset zameraný výlučne na tvar nebol doteraz publikovaný, rozhodli sme sa naše experimentálne dáta verejne sprístupniť. Táto práca už bola čiastočne publikovaná v [235].

V našom experimente sme 73 študentom zobrazili siluety abstraktných a reálnych objektov na jednotnom pozadí.

Na vyhodnotenie nápadnosti tvaru a kontúr sme navrhli niekoľko modelov, ktoré sme overili na fixáciách z experimentu. Prvá skupina modelov vypočítava nápadnosť jednotlivým objektom bez ohľadu na ich okolie. Vyššiu nápadnosť priraďuje väčším, asymetrickým a nepravidelným objektom, ktoré sú reprezentované jednoducho, prostredníctvom obsahu, obvodu, ekvivalentného priemeru, výstrednosti, pomeru strán, rozlohy voči ich ohraničujúcemu obdĺžniku, pravouhlosti, celistvosti a kruhovitosti. Druhá skupina modelov hľadá objekty odlišné od ostatných. Na nájdenie jedinečných objektov sú použité jednoduché charakteristiky využívané prvou skupinou modelov, Hausdorffova vzdialenosť, kontext tvaru, vzdialenosť od stredu objektu v priestorovej a frekvenčnej doméne a momenty [222, 11]. Na rozdiel od predchádzajúcich modelov vypočítava posledná skupina nápadnosť každej kontúry. Tieto modely sú založené na predpoklade, že nápadná kontúra sa odlišuje od svojho okolia. Prvý z kontúrových modelov (CSCD) vychádza z modelu Ittiho a kol. [96]. Preto zo vzdialenosti od stredu objektu vytvorí Gaussovu pyramídu, v ktorej porovnáva hrubšie a jemnejšie vrstvy. Druhý kontúrový model (SRCD) zas využíva vzdialenosť od stredu objektu vo frekvenčnej doméne. Tento model vychádza z modelu autorov Hou a Zhang [86]. Nápadnosť je preto definovaná pomocou tzv. spektrálnych reziduí. Okrem vlastných modelov sme vyhodnotili tiež model nápadnosti tvaru založený na Jaccardovom indexe podobnosti [37] ako aj 2 štandardné modely nápadnosti [96, 80], ktoré počítajú nápadnosť intenzity a orientácie, ktoré sa môžu podieľať na vnímaní tvarov objektov.

Z experimentu sme zistili, že pozornosť smeruje k väčším objektom, ale jasný trend vo fixovaní asymetrických a komplexnejších objektov sme nenašli. Ďalej sme zistili, že globálny kontrast tvaru objektov zohráva len zanedbateľnú úlohu v pozornosti. Naopak, kontrasty v kontúrach objektov mali silný vplyv na pozornosť. Náš kontúrový model SRCD dosiahol signifikantne lepšie výsledky než ostatné modely. Okrem toho na pozornosť počas experimentu vplývali aj vysoko-úrovňové faktory. Participanti často fixovali siluety patriace živým objektom, predovšetkým ich hlavy.

Vizuálna pozornosť ovplyvnená egocentrickou hĺbkou

Napriek tomu, že pozornosť ovplyvňuje aj hĺbka, modely nápadnosti ju využívajú len zriedka. Na rozdiel od predchádzajúcich štúdií, ktoré skúmali pozornosť na 2D a 3D obrazoch [98, 125, 202], náš experiment, ktorého fixačné dáta sú verejne dostupné, analyzuje pozornosť v reálnom prostredí. Pretože binokulárne videnie napomáha rozlišovať hĺbku [5, 164], naším cieľom bolo zistiť, či sa nápadnosť stereoskopickej reálnej hĺbky odlišuje od obrazovej hĺbky.

V našom experimente sme zaznamenávali pohľad 28 študentov v miestnosti so zavesenými identickými guľami, ktoré boli umiestnené v rôznej hĺbke, až do vzdialenosti 4 m od pozorovateľa.

Na rozdiel od sledovania obrazov, kde boli najviac fixované najbližšie objekty, pohľad našich participantov smeroval rýchlejšie a častejšie k vzdialenejším objektom, avšak tento vzťah nie je lineárny. Napriek tomu, že participanti uprednostňovali vzdialenejšie objekty, veríme že, takýto vzťah platí len pre kratšie vzdialenosti ako v našom experimente. Pri väčších vzdialenostiach očakávame, že vzdialené objekty budú natoľko malé, že ich pozornosť bude skôr ignorovať. Okrem toho, pohľad participantov smeroval k objektom, ktoré sú vzdialené od ostatných objektov, tento vplyv bol ale tiež nelineárny. Kvôli týmto zisteniam odporúčame na modelovanie pozornosti v reálnom prostredí špecializované modely.

Egocentrická vizuálna pozornosť ovplyvnená statickými faktormi

Naše experimenty analyzovali vplyv jednotlivých statických stimulov separátne. Naším cieľom je preto zistiť, ako tieto stimuly súperia o ľudskú pozornosť v reálnom prostredí. Na rozdiel od predchádzajúcich prác, sme okrem samotného pohľadu participantov zaznamenávali aj hĺbku scény [215, 193]. Náš experiment už bol čiastočne publikovaný v [233].

V našom experimente sme zaznamenávali pohľad 6 študentov, ktorí voľne preskúmavali laboratórium, v ktorom sa nachádzali iba statické objekty. Okrem toho sme použili zariadenie Kinect, aby sme určili, ako boli od pozorovateľa vzdialené jednotlivé objekty.

V experimente sme sa zamerali na vplyv statických faktorov ovplyvňujúcich pozornosť ako intenzita, farba, orientácia, hĺbka, tvar a kontúry objektov a centrálne vizuálne pole pozorovateľa, ktorých nápadnosť sme predpovedali prostredníctvom existujúcich ako aj vlastných modelov. Na predikciu nápadnosti intenzity, farby a orientácie sme použili model od Ittiho a kol. [96], model od Harela a kol. [80] a vlastný superpixelový model [234], ktorý dáva do korelácie superpixelové histogramy intenzity a orientácie na jemnejších a hrubších vrstvách Gaussovej pyramídy (najvyššiu nápadnosť majú superpixely bez korelácie) alebo porovnáva ich farebné vzdialenosti. Na modelovanie hĺbkovej nápadnosti sme použili jednoduchý

model, ktorý priraďuje nápadnosť hĺbkam lineárne, pričom najbližšie vzdialenosti sú najnápadnejšie. Ďalej sme využili náš experiment s hĺbkou a pomocou získaných fixačných dát sme aproximovali funkciu nápadnosti. Takýto model považoval za najnápadnejšie najvzdialenejšie oblasti. Okrem toho sme predikovali nápadnosť globálneho hĺbkového kontrastu pomocou jednoduchého modelu, ktorý priraďoval nápadnosť kontrastu lineárne a modelu, ktorý túto nápadnosť priraďuje nelineárne, na základe fixácií získaných z nášho experimentu s hĺbkou. Nakoniec sme navrhli model lokálneho globálneho kontrastu, ktorý funguje podobne ako náš superpixelový model v prípade predikcie nápadnosti intenzity [234]. Na modelovanie nápadnosti tvaru sme použili jednoduché vlastnosti objektov – obvod a ekvivalentný priemer. Tieto modely priraďujú vyššiu nápadnosť väčším objektom a objektom, ktoré sa v týchto vlastnostiach od ostatných odlišujú. Ďalej sme predikovali nápadnosť kontúr na základe vzdialenosti od stredu objektu pomocou modelov CSCD a SRCD. Nakoniec sme modelovali centrálne vizuálne pole ako Gaussova funkcia umiestnená uprostred obrazu.

Analyzovaním fixácii sme zistili, že najväčší vplyv na pozornosť majú vo všeobecnosti intenzita, farba a orientácia. Okrem toho, ale výsledky naznačujú silnú individualitu vo fungovaní pozornosti našich participantov. Ich pozornosť mohla byť pravdepodobne výrazne ovplyvnená vysoko-úrovňovými faktormi ako identifikácia objektov a prekvapenie. Intenzita bola najlepšie predikovaná naším superpixelovým modelom [234], zatiaľ čo farbu a orientáciu najlepšie predikoval model Harela a kol. [80]. Naopak, modely založené na vzdialenosti objektov od pozorovateľa a globálnom hĺbkovom kontraste mali na pozornosť len zanedbateľný vplyv. Napriek tomu, naša experimentálna funkcia nápadnosti hĺbky priniesla zlepšenie predikcie u 4 participantov. Všetky hĺbkové modely sú ale výrazne prekonané naším modelom založenom na lokálnom hĺbkovom kontraste. Nápadnosť tvaru bola úspešne predpovedaná len našimi kontúrovými modelmi. Okrem toho sme zistili, že pozornosť participantov smerovala do stredu ich vizuálneho poľa. Vplyv týchto príznakov sa ale v čase menil. Výsledky naznačujú, že najskôr ovplyvňujú pozornosť najzákladnejšie príznaky ako intenzita, farba a orientácia na rozdiel od tvaru, ktorého nápadnosť neskôr mierne zosilňuje.

Vizuálna pozornosť ovplyvnená emóciami

Psychologické štúdie potvrdili prepojenie medzi emocionálnymi stimulmi a vizuálnou pozornosťou, na rozdiel od vplyvu nálady na vizuálne spracovanie emociálne neutrálnych stimulov, ktorému sa venovali len okrajovo. Na rozdiel od predchádzajúcich prác náš experiment skúma, či zdola nahor nízko-úrovňová nápadnosť ovplyvňuje pozitívna nálada. I keď existujú modely nápadnosti určené pre emocionálne stimuly, klasické modely nápadnosti sa nikdy nevyhodnocovali voči pozorovateľom v určitej emócii.

Náš experiment skúma pozornosť 10 ľudí, ktorým bola vyvolaná pozitívna alebo neutrálna emócia prostredníctvom vlastných spomienok. V prípade pozitívnej nálady si participanti mali spomenúť na ich šťastnú udalosť z ich života, zatiaľ čo pri neutrálnej nálade mali opísať cestu do laboratória, kde bola emócia vyvolaná. Participanti sa potom presunuli do druhej miestnosti, kde im bola zobrazená sériá prirodzených a umelých obrazov voľne, s cieľom zapamätať si ich obsah alebo čo najrýchlejšie nájsť cieľový objekt. V prípade vizuálneho hľadania mali participanti nájsť na zobrazenej scéne jediný objekt, ktorého vzhľad vyhovuje zadaniu, jeden z viacerých možných cieľov, ktorých vzhľad odpovedá zadaniu alebo nájsť jedinečný objekt. Na modelovanie zdola nahor nápadnosti sme použili štandardný model [96].

Napriek tomu, že dotazník PANAS ukázal podobné výsledky v hodnotení ich nálady, medzi oboma skupinami emócií sme našli určité rozdiely. Pri vyhodnocovaní ich fixácii sme očakávali rozšírenie vizuálnej pozornosti vplyvom pozitívnej nápady na základe teórie rozširovania a budovania [62]. Avšak čas, za ktorý úlohu vyriešili, skôr naznačuje úplný opak – pozitívna nálada rozptýlila participantov od hľadania cieľa. I keď sme nenašli zvýšenú podobnosť fixácií pre žiadnu emóciu, rozdiely sme našli vo vplyve zdola nahor nápadnosti. Pri riešení úloh nájsť jedinečný objekt a zapamätať si scénu sme našli zvýšený vplyv nápadnosti v neutrálnej nálade. Naopak pri voľnom prezeraní scény bol tento vplyv dominantnejší v pozitívnej nálade. Toto zistenie by sa mohlo vysvetliť nižšou angažovanosťou participantov v prípade prezeranie scén bez žiadnej úlohy, ktorej vplyv na pozornosť naznačili už aj predchádzajúce štúdie [186, 162, 99]. Preto predpokladáme, že rozširovanie pozornosti pozitívnou náladou nastáva len v spojení s nízkou úrovňou angažovanosti, príp. s nižšou náročnosťou úloh.

Vizuálna pozornosť ovplyvnená cieľmi v informačných vizualizáciách

To, ako sa človek pozerá na vizualizácie ovplyvňujú nápadné stimuly, ale aj znalosti, záujem a úlohy. Napriek tomu, že sa v poslednom období modely nápadnosti používajú na predikciu pozornosti vo vizualizáciách pri exploratívnej analýze, o vplyve zdola nahor nápadnosti pri konfirmačnej analýze vieme stále málo. Na rozdiel od predchádzajúcich experimentov, náš experiment skúma pozornosť vo vizualizáciách počas riešenia 3 jednoduchých analytických úloh. Tento experiment už bol publikovaný v [236].

V našom experimente analyzujeme pozornosť 47 študentov, ktorí mali za úlohu vyriešiť úlohy v rôznych vizualizáciách čo najrýchlejšie a najsprávnejšie. Vizualizácie sme prebrali z datasetu MASSVIS [21], ktoré boli pôvodne použité v experimente, v ktorom mali participanti za úlohu zapamätať si ich obsah. Z tohto datasetu sme použili stĺpcové grafy, geografické mapy, plošné grafy, bodové grafy, tabuľky a čiarové grafy. Pre každú použitú vizualizáciu sme navrhli 3 úlohy ako získanie hodnoty určitého dátového elementu, filtrovanie elementov na základe zadaného kritéria a nájdenie extrémnej hodnoty. Okrem toho sme navrhli pre každý typ úlohy optimálnu stratégiu pre jej vyriešenie. V prípade získania hodnoty sme predpokladali, že participanti najskôr hľadajú názov cieľového elementu, ktorý si potom spoja s odpovedajúcim elementom a nakoniec odčítajú hodnotu jeho atribútu, ktorý majú za úlohu zistiť. Naopak v prípade filtrovania a nájdenia extrému sme očakávali, že participanti najprv nájdu vyhovujúce hodnoty, potom si ich spoja s elementami, ktorým hodnoty patria a na záver prečítajú názov týchto cieľových elementov. V prípade hľadania extrémnej hodnoty sa môže v prípade určitých vizualizácii odčítanie samotnej hodnoty preskočiť a rovno fixovať samotný cieľový element (napr. v bodových grafoch kde sú elementy zoradené podľa atribútu, ktorého extrém sa hľadá). Takto získané fixácie počas konfirmačnej analýzy sme napokon porovnali s viac exploratívnou analýzou počas zapamätania [20]. Na predikciu nápadnosti sme použili 12 existujúcich modelov vrátane konvolučných neurónových sietí a špecializovaného modelu pre vizualizácie (DVS), ktorý kombinuje štandardnú nápadnosť od Ittiho a kol. [96] a nápadnosť textu.

Zistili sme, že fixácie počas riešenia úloh sú súdržnejšie než pri exploratívnej analýze počas úlohy zapamätania. Okrem toho výsledky ukázali zvýšenú podobnosť fixácií medzi úlohami nájdenia extrému a zapamätania, čo možno vysvetliť tak, že extrémy sú reprezentatívne hodnoty pre zapamätanie. Časy prvého fixovania cieľových častí vizualizácie ukázali, že optimálnu stratégiu participanti použili iba pri riešení úlohy získania hodnoty. Ďalej sme

zistili, že pozornosť participantov najlepšie predikuje špecializovaný model DVS, pričom pri exploratívnej analýze bol výrazne úspešnejší než pri konfirmačnej analýze. Preto možno konštatovať, že pri voľnom prezeraní úloh je vplyv zdola nahor nápadnosti vyšší než pri riešení úloh. Ďalej sme zistili, že nápadnosť [96] objektu neovplyvňuje rýchlosť ich fixovania. Extrémne hodnoty ale nie sú ani nápadnejšie podľa výpočtového modelu [96], ani rýchlejšie fixované než cieľové objekty zvyšných dvoch úlohách. Na základe výsledkov odporúčame pre vizualizácie špecializované modely, ktoré budú pracovať na úrovni objektov a identifikovať jednotlivé prvky vizualizácie a ich vzťahy, napr. legendy, osy a popisy jednotlivých elementov.

Zhrnutie

Táto práca pojednáva o viacerých faktoroch, ktoré ovplyvňujú vizuálnu pozornosť (stimulmi-riadené a aj cieľmi-riadené faktory). Vykonali sme vlastné experimenty so sledovaním pohľadu, navrhli výpočtové modely pozornosti a diskutovali o fungovaní vizuálnej pozornosti. Modelovanie pozornosti potrebuje špecializované datasety na zlepšenie predikcie nápadnosti. Preto sme naše fixačné datasety verejne sprístupnili.

Naše experimenty skúmali tieto faktory separátne. Väčšina z nich ukázala obrovské rozdiely v ich vplyve na vizuálnu pozornosť a jej výkon, predovšetkým v prirodzenom prostredí z pohľadu prvej osoby.

Skúmali sme nápadné 2D príznaky ako farba a tvar na vlastných umelých obrazoch. Zistili sme, že farebné kontrasty v LAB farebnom priestore neovplyvňujú pozornosť rovnako a preto by sa pri modelovaní pozornosť mali zvážiť aj vysoko-úrovňové aspekty farby. Asociácie farieb k život ohrozujúcim situáciám, ako napr. červená indikujúca nebezpečenstvo, však mali len zanedbateľný vplyv na pozornosť. Ďalej sme ukázali, že pohľad smeruje ku kontrastom v kontúrach objektov. Túto nápadnosť modelujú s vysokou presnosťou spektrálne reziduá vo vzdialenosti od stredu objektu. Okrem toho sú častejšie fixované väčšie objekty.

Okrem toho sme poukázali na to, že egocentrická pozornosť v reálnom prostredí sa odlišuje od prezerania obrazov, preto sú pre jej predikciu potrebné špecializované modely nápadnosti. Pozorovali sme, že binokulárne videnie smeruje k objektom vzdialenejším od pozorovateľa a od okolia, na rozdiel od hĺbky v obrazoch, kde sú často fixované najbližšie objekty. Aj keď pohybujúce sa a prekvapivé stimuly výrazne ovplyvňujú pozornosť, zistili sme, že statické stimuly dominujú nad tými dynamickými. Naše experimenty odhalili, že najvýraznejší vplyv na pozornosť majú kontrasty v intenzite, farbe a orientácii. Rovnako sme ukázali, že egocentrická pozornosť smeruje do stredu vizuálneho poľa pozorovateľa.

Okrem nízko-úrovňových vizuálnych príznakov môže pozornosť ovplyvňovať aj momentálny emocionálny stav. Pozitívna emócia vyvoláva silnejší vplyv nápadnosti pri voľnom prezeraní emocionálne neutrálnych stimulov. Avšak opačný efekt sme pozorovali pri riešení úloh. Taktiež sme zistili, že ľudia v pozitívnej nálade môžu riešiť úlohy pomalšie, preto si myslíme, že ľudí skôr rozptyľuje od úloh.

Výpočtové modely nápadnosti sa využívajú v mnohých oblastiach informatiky, predovšetkým pre prirodzené obrazy, kde neurónové siete úspešne znížili rozdiel medzi predpovedanou nápadnosťou a ľudskými fixáciami. Tieto modely sa taktiež používajú vo vizualizáciách, napr. ako metrika kvality. Kvôli rozdielom medzi prirodzenými a umelými obrazmi sa náš

posledný výskum zameral na informačné vizualizácie a konfirmačnú analýzu. Tieto výsledky ukázali, že vplyv faktorov riadených cieľmi výrazne stúpa počas rôznych úloh hľadania. Pre vizualizácie odporúčame špecializované modely nápadnosti. Na rozdiel od prirodzených obrazov výrazne vplýva na pozornosť nápadnosť textu.

V budúcnosti by sa hore uvedené zistenia z experimentov mali zlúčiť do výpočtového modelu, ktorý by tak mohol spoľahlivo predpovedať pozornosť v prirodzených obrazoch a špecializovaných doménach, ako napr. informačné vizualizácie a medicínske snímky. Kvôli rozdielnosti v spracovávaní vizuálnej informácie by riešením mohlo byť použitie hlbokých neurónových sietí, ktoré by sa mohli naučiť preferencie vo fixovaní konkrétneho človeka pri riešení určitej úlohy.

Appendix C

Publications of the Author with Internal Categorization and Relevant Citations

The work of this thesis is based on the following publications:

Journal publications – publication category ADC

POLATSEK, Patrik – BENEŠOVÁ, Vanda – PALETTA, Lucas – PERKO, R. Novelty-based Spatiotemporal Saliency Detection for Prediction of Gaze in Egocentric Video. In *IEEE Signal Processing Letters*. Vol. 23, iss. 3 (2016), pp. 394-398. ISSN 1070-9908 (2016: 2.528 - IF, 2 - JCR Best Q, 0.798 - SJR, Q1 - SJR Best Q). In databases: WOS: 000372320000001; SCOPUS: 2-s2.0-84962301435.
Citations:

1. ARDESHIR, Shervin – BORJI, Ali. Ego2top: Matching viewers in egocentric and top-view videos. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016-01-01, 9909, pp. 253-268. ISSN 03029743., Registered in: SCOPUS, WOS
2. TTANG, Zhenhua – LUO, Yadan – ZHANG, Rui – JIANG, Hongbo. Motion saliency detection for compressed videos. In Journal of Electronic Imaging, 2017-09-01, 26, 5, pp. ISSN 10179909., Registered in: SCOPUS
3. GORJI, Siavash – CLARK, James J. Going from Image to Video Saliency: Augmenting Image Salience with Dynamic Attentional Push. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018-12-14, pp. 7501-7511. ISSN 10636919., Registered in: SCOPUS, WOS
4. TAVAKOLI, Hamed R. – RAHTU, Esa – KANNALA, Juho – BORJI, Ali. Digging deeper into egocentric gaze prediction. In Proceedings 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019, 2019-03-04, pp. 273-282., Registered in: SCOPUS

POLATSEK, Patrik – WALDNER, Manuela – VIOLA, Ivan – KAPEC, Peter – BENEŠOVÁ, Vanda. Exploring visual attention and saliency modeling for task-based visual analysis. In *Computers and Graphics*. Vol. 72, (2018), pp. 26-38. ISSN 0097-8493 (2017: 1.200 - IF, 3 - JCR Best Q, 0.355 - SJR, Q2 - SJR Best Q). In databases: SCOPUS: 2-s2.0-85042471777;

WOS: 000431158300005.
Citations:

1. BEHRISCH, M. – BLUMENSCHEIN, M. – KIM, N. W. – SHAO, L. – EL-ASSADY, M. – FUCHS, J. – SEEBACHER, D. – DIEHL, A. – BRANDES, U. – PFISTER, H. – SCHRECK, T. – WEISKOPF, D. – KEIM, D. A. Quality Metrics for Information Visualization. In COMPUTER GRAPHICS FORUM, 2018, vol. 37, no. 3, pp. 625-662. ISSN 0167-7055., Registered in: WOS, SCOPUS

Conference publications – publication category AFC

OLEŠOVÁ, Veronika – BENEŠOVÁ, Vanda – POLATSEK, Patrik. Visual attention in egocentric field-of-view using RGB-D data. In *Proceedings of Ninth International Conference on Machine Vision (ICMV 2016)*, 18th November, 2016, Nice, France, iss. 1: SPIE - The International Society for Optical Engineering, 2017, pp. 1-9. ISBN 978-151061131-3. In databases: WOS: 000410664800028; SCOPUS: 2-s2.0-85029924145.
Citations:

1. AKSIT, Kaan – CHAKRAVARTHULA, Praneeth – RATHINAVEL, Kishore – JEONG, Youngmo – ALBERT, Rachel – FUCHS, Henry – LUEBKE, David. Manufacturing Application-Driven Foveated Near-Eye Displays. In IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 2019, vol. 25, no. 5, pp. 1928-1939. ISSN 1077-2626., Registered in: WOS, SCOPUS

POLATSEK, Patrik – JAKAB, Marek – BENEŠOVÁ, Vanda – KUŽMA, Matej. Computational models of shape saliency. In *Proceedings SPIE 11041, Eleventh International Conference on Machine Vision (ICMV 2018)*, 2018, Munich, Germany, iss. 1 Bellingham: SPIE - The International Society for Optical Engineering, 2019, ISBN 9781510627482. In databases: DOI: 10.1117/12.2522779.

Conference publications – publication category AFD

POLATSEK, Patrik – BENEŠOVÁ, Vanda. Bottom-up saliency model generation using superpixels. In *SCCG 2015. Proceedings of the 31st Spring Conference on Computer Graphics*, iss. 1 New York: ACM, 2015, pp. 121-129. ISBN 978-1-4503-3693-2. In databases: SCOPUS: 2-s2.0-84963525793; WOS: 000380609300018.