SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
Faculty of Informatics and Information Technologies

FIIT-5208-72202

Bc. Miroslav Laco

GENERATING A SALIENCY MAP WITH FOCUS ON DIFFERENT ASPECTS OF HUMAN VISUAL ATTENTION

Master thesis

Degree course:	Information systems
Field of study:	9.2.6 Information systems
Training workplace:	Institute of Computer Engineering and Applied Informatics
Supervisor:	doc. Ing. Vanda Benešová, PhD.

May 2018

Anotácia

Slovenská technická univerzita v Bratislave

FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

Študijný program: Informačné systémy

Autor: Bc. Miroslav Laco

Diplomová práca: Model predikcie ľudskej vizuálnej pozornosti so zameraním na rôzne aspekty vizuálneho vnemu človeka

Vedúci diplomovej práce: doc. Ing. Vanda Benešová, PhD.

máj 2018

Modelovanie ľudskej vizuálnej pozornosti je posledné roky predmetom rozsiahleho výskumu. V prvej časti diplomovej práce analyzujeme súčasný stav v oblasti modelovania vizuálnej pozornosti. Následne predstavujeme nový prístup k získavaniu dát o ľudskej vizuálnej pozornosti z egocentrického pohľadu v reálnom prostredí, pričom sa opierame o predošlé práce.

V práci opisujeme nový a kompletný návrh pre experimenty s ľudskou vizuálnou pozornosťou v laboratóriu. Používame špecifický hardvér a navrhujeme algoritmy, využívajúce metódy počítačového videnia, ktoré napomáhajú k dosiahnutiu navrhovaného cieľa. Návrh sme implementovali a viedli sme rozsiahle experimenty s ľudskou vizuálnou pozornosťou, ktorých výsledkom je nový súbor dát pre skúmanie ľudskej vizuálnej pozornosti z egocentrického pohľadu.

Jedným z najväčších prínosov nášho návrhu a nového súboru dát je možnosť skúmať aspekty, vplývajúce na ľudskú vizuálnu pozornosť, ktoré nebolo možné skúmať nikdy predtým. Na základe predošlých prác sme súbor dát využili na výskum vplyvu hĺbky scény reálneho sveta na vizuálnu výraznosť objektov v nej. Tvrdíme, že vplyv hĺbky na vizuálnu výraznosť objektov je aplikovateľný na súčasné modely vizuálnej pozornosti ako koeficient vizuálnej výraznosti. Výsledky nášho výskumu sme aplikovali na existujúci model vizuálnej výraznosti, vyhodnotili sme výsledky nového modelu a zhodnotili sme možné vylepšenia.

Annotation

Slovak University of Technology Bratislava

FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Degree course: Information systems

Author: Bc. Miroslav Laco

Master's thesis: Generating a saliency map with focus on different aspects of human visual attention

Supervisor: doc. Ing. Vanda Benešová, PhD.

2018, May

Visual attention modelling is under extensive research throughout the past years. We analyse the current state-of-the-art in the visual attention modelling in the analytical part of this thesis. We propose a novel method to conduct user studies for research of human visual attention in real world environments from the egocentric perspective of view, building upon state-of-the-art in the visual attention modelling.

We introduce a novel and complete method proposal for the user studies setup in a laboratory. To meet our specified goals, we use various hardware equipment and introduce our own algorithms and procedures based on the principles of image processing and computer vision. We created a novel dataset for studying human visual attention in real environments from the egocentric perspective of view during the extensive user studies following our proposed method.

One of the biggest assets of the proposed method and the created dataset is the possibility to study aspects affecting visual attention that were not possible to study before. Based on the previous work in the field, we decided to conduct a research on the depth influence (distance between the observer and the observed object) on visual attention in real environments using the novel dataset. We claim that the aspect of depth influence on human visual attention can be applied on existing visual attention models as a saliency coefficient. We apply the results of our research on an existing saliency model, summarize up the results and conclude future possible improvements.

Declaration of honour

I honestly declare, that I wrote this thesis independently under supervision of doc. Ing. Vanda Benešová, PhD. with cited bibliography.

May 2018 in Bratislava .....

Miroslav Laco

Acknowledgement

I would like to express my special thanks of gratitude to my supervisor doc. Ing. Vanda Benešová, PhD. for her willingness, continual support, professional supervision, valuable advices and helpful ideas during the work on this thesis. Secondly, I would like to thank Ing. Peter Kapec, PhD. who helped me with human resources to make my research more relevant. Lastly, I would like to express my sincere thanks to my relatives and friends for their generous support.

Contents

1 Introduction

Sight is one of the five human senses through which we can explore our surroundings. Our brain has certain capacity for processing signals that has their origin in our senses. Human visual perception takes approximately 80% of this capacity. People, however, cannot process all visual signals perceived by their sight, despite of the high amount of brain capacity denoted for visual perception. Our visual perception is overwhelmed with visual stimuli from the objects all around us. That is a reason why visual perception system of humans consist of selection mechanism applied on perceived stimuli and notion of relevance (Borji – Itti, 2013). It is known that visual perception systems apply serial computational strategy despite of the fact that the process appear parallel to us. The parallelism is only the illusion of fast sequential processing of multiple tasks, just like we know from the CPU processing in the computer science. Thus, a particular location on the scene is selected for processing once at a time and its surroundings are suppressed and referenced as a fringe (Itti, 2000; Polatsek, 2015). While visual perception's anatomical structure is familiar to us these days, the focus of research in visual perception is visual attention and underlying computational mechanisms (Borji – Itti, 2013). Visual attention is a process that helps us to decide where to fix our attention and which visual stimuli to process first by determining the salient regions (Goldstein, 2010).

Scientific research related to the visual attention is interdisciplinary and involves work of psychologists (Rensink, 1997; Simons – Chabris, 1999), neurobiologists (Treue, 2001; Kastner – Ungerleider, 2000) and computer vision scientists who take the benefits from the research in medical and psychological research fields. Computer vision scientists focus mainly on problems of mathematical descriptions and complexity of visual attention and its applications in real time (Borji – Itti, 2013). The ultimate aim of their research is to determine and predict visual attention by models of visual attention. To build such a model, one has to analyse human eye fixations and movements on the scenes- the gaze information (Goldstein, 2010). We can use various equipment to gather gaze data during the user studies focused on the visual attention research and we can use many approaches to analyse and conclude them. The result of applying visual attention model on a given scene is visual attention map (or a saliency map) which predicts eye fixations on scene (Borji – Itti, 2013).

1.1 Motivation

To analyse the human visual attention and to introduce its model is hard and complex task and is subject of research for a few decades. Nowadays, we know certain amount of aspects that influence visual attention. However, there is a lot of work left to reveal more of them and to determine their exact impact on models of visual attention. Moreover, this task is not so easy because every human is a unique individual and therefore visual attention model of every human is unique. Our task is, therefore, to look for some common patterns that apply for majority of us. Among well explored aspects of visual attention belong static visual stimuli such as colour, orientation or contrast. Based on the knowledge from the last years we know that these stimuli are not sufficient for modelling our visual attention as a whole. We should take into account more aspects which affect our visual attention such as distance of the objects

on the scene and many more unexplored ones. The unexplored aspects influencing visual attention require even more extensive user studies and novel approaches for studying the visual attention. The knowledge from the research in the field of visual attention modelling is getting more and more popular because of large scope of its applications.

Biologically inspired computer vision is widely used in robotics where further development of robotic active vision and human-robot interaction crucially depends on the principles of visual attention modelling. There is a suggestion that people's trust in the intelligent robots is related to their attention patterns (Nagai et al., 2008). Interaction between human and robot, as well as between human and human, may be disturbed by number of factors. These distractions prevent human from building trustful and positive relationship with interacting other. Therefore, there is a strong interest in eliminating such distractions from robot's behaviour and developing visual attention of intelligent robots as close to humans as possible (Nagai et al., 2008).

Among popular applications of visual attention models belong computer vision and graphics. We know several algorithms that can break down image into segments and recognize objects from them. This is task which can be done with benefit using the models of visual attention (Mitri et al., 2005). More complex task, however, is to prioritize objects on image from the least important to the most important ones for people. This way, the computer can decide which of them to highlight, show to user or make a thumbnail from. This involves principles of human visual attention applied on images (Marchesotti et al., 2009). Similar knowledge can be used in image and video compression algorithms when the task is to maintain details of most salient objects while discarding them from others with lower saliency (Itti, 2004).

Advertising is category of application of visual attention models in contrast with technical-oriented applications mentioned above. Advertising goals in all kinds of media is to send clear message to potential future buyers in the shortest time possible. This is the reason why advertisement experts involve visual attention scientists in their teams to improve effectivity of their advertisement campaigns (Liu et al., 2008).

2 Human vision and visual attention

Human visual system is a complex structure responsible for receiving and evaluating visual information from the world around us. The amount of visual information from our surroundings is so large that our brain must process signals from eyes selectively using principles of visual selectivity and priority. Definition of visual attention is not so clear even until nowadays- mainly because of certain abstractness of this phenomenon. In order to find one of the first written definition we have to move a few decades backwards where psychologist W. James said: *" Everyone knows what attention is. It is the taking possession of the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence."* (James, 1890). We will look more in detail on the visual attention as a complex process to understand it before moving to the visual attention modelling research.

2.1 Human vision

Human vision from the biological perspective is a sensual system consisting of two receptor subsystems located in eyes, nerves that transport perceived signals and visual cortex in brain responsible for evaluating incoming sensual messages. Vision system is made up of many subsystems specialized on identification of contrast, shape, motion, depth, colour and many more characteristics of objects (Dobeš, 2005).

Light is a term for visible electromagnetic radiation which is in range of 400-700 nanometres. Light that enters the eye is transformed on retina into electric impulses by cells responsible for conversion, so called ganglion neurons (Dobeš, 2005). Another important cell found on retina are rods and cones. These are light-sensitive photoreceptors with different specialization. Cones are sensitive on details and provides us colour vision while rods are sensitive to radiation of light with small energy and helps us to see in dark conditions and perceive outlines of objects without colour information. Rods can be found on retina with significantly higher percentage- there are approximately 20 times more rods than cones there. Cones are concentrated mainly in the fovea which is small part on retina with best visual perception capabilities (Polatsek, 2015; Dobeš, 2005).

Electric stimuli of visual information are transported to the visual cortex in the brain by the optical nerves. Average ratio of receptor cells number in eye and nerve cells number in optical nerve is about 130:1 which leads us to conclusion that considerable part of visual stimuli pre-processing is happening directly in the eye (Dobeš, 2005).

Motion of the eyes is possible using six eye muscles (for both eyes twelve) which are controlled in the frontal lobe of the brain in the three principal areas: the frontal eye field, the supplementary eye field and the dorsolateral prefrontal cortex (Pierrot-Deseilligny et al., 2004). Humans eye movement scan paths typically consists of alternating fixations and saccades. The fixations represent information gathering sequences around an interest region and saccades indicate the transitions between fixations (Karthikeyan, 2013). Frontal eye field is strongly connected with the saccades. It is involved in the preparation and triggering of saccades. These can be of five types based on the target location and behaviour:

- Intentional, visually guided saccade (towards a target already present)

- Predictive saccade (towards a target not yet present)
- Memory-guided saccade (target no longer visible)
- Anti-saccade (towards opposite direction than target)
- Pro-saccade (reflexive, visually guided saccade)

Pro-saccades are mainly triggered earlier than in frontal eye field, namely in parietal eye field and are triggered towards a suddenly appearing peripheral target as a reflex reaction (Pierrot-Deseilligny et al., 2004). Movement of the eyes is continual never-ending process where saccades and fixations are alternated (Dobeš, 2005). This is because of the anatomy of receptor cells on our retina. They are working in such way that they need change every small period of time to deliver visual stimuli to visual nerve and by visual nerve to visual cortex. You can prove this theory by looking constantly on black dot on a blank paper. After few seconds the dot will disappear just as if you were blind (Dobeš, 2005).

Visual cortex in the brain consists of a few areas as shown on Figure 2-1. The smallest anatomical parts of visual cortex are visual nerve cells that are working on their own in the lower parts of the brain or are grouped together to form neural networks in the higher parts responsible for complex tasks. Similarly, as we know from the theory of computer neural networks, visual nerve cells do have an input and on-off-like output.

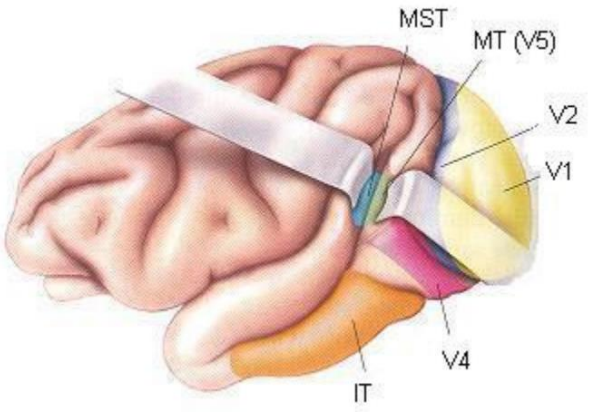


Figure 2-1: Visual cortex parts from the macaque's brain. Human brain is very similar in the means of the location of visual cortex and its parts (Dobeš, 2005). The cortex parts are referenced further in the subchapter.

Primary visual cortex (V1) is the first part of a brain that processes incoming neural stimuli from the visual nerves. It consists from simple neurons and complex neurons. Simple neurons detect simple geometric shapes with their alignments such as one neuron for horizontally aligned rectangle and another for rectangle at 45-degree angle (Dobeš, 2005). Complex neurons are evaluating information from the simple neurons to build up more complex information which can be movement of a rectangle as an example (Connors et al., 2007). Important fact is that object boundaries are most of the time enough for the brain to identify it (Dobeš, 2005).

Secondary visual cortex (V2) is responsible for detection of contours of more complex objects and V4 area detects a shape of the objects. Neurons in this area are sensitive to the colour and light, too. One can also find special neurons which can detect important objects for humans because of the human evolution- faces, silhouettes, hands, animals, and many more. The most complex objects have to be further examined and evaluated in higher areas of the middle temporal visual area. All these processes

are not taking place only in one direction- for example from V1 area to V4 and further- but they happen in all directions as needed to process every single stimulus (Dobeš, 2005; Connors et al., 2007).

2.2 Visual attention

Despite of the fact that definition of the attention as a term is an easy task, we can find out that the visual attention is a phenomenon covering all the factors that influence our selection mechanisms, whether they are stimulus driven (bottom-up) or expectations driven (top-down) (Borji – Itti, 2013).

The meaning of visual saliency is often misunderstood as many people think that this term is visual attention synonym. More accurate definition of the saliency says that it just describes characteristics of some parts of scene that stands out in comparison to others because of physical properties such as colour, brightness, contrast, or orientation (Borji – Itti, 2013). Capturing attention by stimulus salience is a bottom-up process as it depends solely on the pattern of stimulation falling on the visual receptors (Goldstein, 2010). Therefore, visual saliency has strong connectivity with context of bottom-up visual attention computations (Koch – Ullman, 1985; Itti et al., 1998). Saliency is making up only one group of aspects that affect our visual attention. Among the others belong top-down stimuli and many more aspects such as face or human body presence on the scene.

2.2.1 Bottom-up processing

Visual stimuli that automatically and involuntarily attract our attention are called bottom-up (scene- or context-driven) stimuli (Borji – Itti, 2013; Itti – Koch, 2001). The higher the visual saliency the larger priority part of the scene has when undergoing bottom-up processing. Key feature of the bottom-app saliency is that part of the scene or object should be standing out in comparison to its neighbours in order to be salient. Typical example of a highly salient bottom-up stimuli originator is the emergency telephone located on sides of highways. These objects stand out from the context of the scene by its shape, colour and often by high contrast due to the reflectance of the material. These all along with stimuli such as depth, unexpected motion, flickering and others are making up the palette of bottom-up stimuli affecting our visual attention pre-attentively.

Pre-attentive influence of perception means an influence that affects the way that we perceive stimuli before we even get the conscious information about the stimuli presence from our brain. Therefore, we cannot change the way the stimuli will influence us in this stage of visual signal processing. An object is in this brief time broken down into the features discussed above (Goldstein, 2010). Working with visual stimuli that affect pre-attentive stage of perception means taking general influence over the perceiving target. This is often applied by scientist and designers that design safety elements in buildings or vehicles, road signs. These can be often misused to inadequately draw our attention to advertisement on web pages or in the real environments.

Characteristics of bottom-up pre-attentive processing is automatics, reflexivity and rapidness- it is very quick process with the rate of 20 to 50 milliseconds for processing an average item (Itti – Koch, 2001).

2.2.2 Top-down processing

This type of visual signal processing brings the observer's knowledge into play and therefore is so called knowledge-based processing (Goldstein, 2010). Top-down processing is affected by our prior experience, memories, skills, observation tasks and many more. This type of processing is taking place in both the pre-attentive and the attentive stage. Therefore, we can partially influence it or take control over it. The top-down processing is significantly slower than the bottom-up one.

Main reason is that the top-down processes are handled in the higher areas of our brain (including frontal lobes) which are connected back to the visual cortex and parts of the visual system that are the originators of perceiving visual stimuli (eyes). The reason for this is that the top-down stimuli often require our more effort to be processed correctly (sometimes even voluntary one). It is also taking its price- average amount of time it takes to process top-down driven visual stimuli is 200 milliseconds and more based on stimuli complexity (Itti – Koch, 2001).

It is important to know that the bottom-up and the top-down processing are creating our visual perception side-by-side and the processes cannot exist without each other. There are some theories bringing the idea that perceiving some special types of stimuli is affected only by the bottom-up processing (Goldstein, 2010), but in general the two types of processing are complementary and are taking place sequentially in the order of bottom-up processing and, subsequently, top-down processing.

3 Theory base for visual attention modelling

Visual attention modelling is undergoing an intensive research nowadays and the models are still not perfect enough. Model of visual attention is usually a result of the research in the field of human visual attention or analysis of some previous research and existing datasets. There are many approaches to create such a model and they greatly differ one from each other while maintaining the same theoretical knowledge base described in the previous chapter.

Imagine yourself looking on a visual scene. At the very first moment of the perception the scene is a static picture with visual stimuli whose perception is influenced by bottom-up and some low impact of top-down processing. Models that consider only this static state of the scene are called spatial models. However, when you look on the scene for a little longer you will realize that the scene, whatsoever static it could seem to be, is in move all the time. Movement ca be of several types regarding self-movement, eye-movement, movement of the whole scene or of the objects on the scene on its own. Therefore, the visual attention is not defined by current state of the scene only but is influenced by accumulated knowledge from the previous time sequences, too (Borji – Itti, 2013). This additional information is most of the time very important and is called temporal information. In general, it is combined with the spatial information to build up the spatio-temporal models of the visual attention.

We move our eyes from three to five times per second in average to align part of the scene we want to percept thoroughly with our fovea (Itti – Koch, 2001). Visual attention that relates to the eye movements and alignment of the fovea with the processed stimuli is known as the overt attention (Polatsek, 2015). This type of visual attention is strongly interconnected with the opposite phenomenon- covert attention. This type of attention describes mental focus onto one of the several visual stimuli on the scene without physically focusing the eye (Borji – Itti, 2013). For better imagination of the covert visual attention, we can think of tracking a person with whom we do not want to come into contact and therefore we avoid focusing our eyes on the person directly. Example of the unconscious covert attention may be walking down the street at night focusing on the target we want to get to while covertly scanning surroundings of the street for a possible threat. Because overt and covert attention are in strong relation with each other, it is important to consider both in the visual attention models. We can easily measure overt attention using the eye-trackers, however, it is a lot harder to measure covert attention. Common approaches to measure covert attention are micro saccades, variances in saccade directions and measuring visual cortex activity through Mind-Machine Interfaces (Hafed – Clark, 2002; Borji – Itti, 2013). Many models omit covert attention influence on overall visual attention due to difficulty of such measures (Borji – Itti, 2013) and due to very little affect that covert attention has on visual attention under certain circumstances.

Methods and mechanisms used for developing saliency models denote the properties and characteristics of the saliency models and often the restrictions for its usability. There is a considerable number of approaches to obtain saliency model summarized in the state-of-the-art summarizing paper of Borji and Itti (Borji – Itti, 2013). General division schema is shown on Figure 3-1. The main difference between them is in the computational mechanisms they adopt. These can vary from biologically-inspired computational mechanisms of cognitive models, probabilistic-inspired computational mechanisms that can take top-down and memory aspects into account with advantage (Bayesian models), probabilistic-inspired mechanisms that involves visual decisions of a viewer with respect to the end of the viewing task (decision theoretic models), combination of probabilistic and graph mechanisms where stimuli for

generating the eye movements are hidden variables with conditional independence denoted by a graph (Borji – Itti, 2013), signal-processing-inspired mechanisms that analyse a scene and compute saliency model rather from the frequency domain than from the spatial domain (spectral analysis models), computer-vision-inspired mechanisms that implement pattern classification approach to accompany the bottom-up saliency with the information about the objects on a scene and possibly their semantics (pattern classification models) and many more. These approaches are mixed and often used together in the visual attention modelling applications.

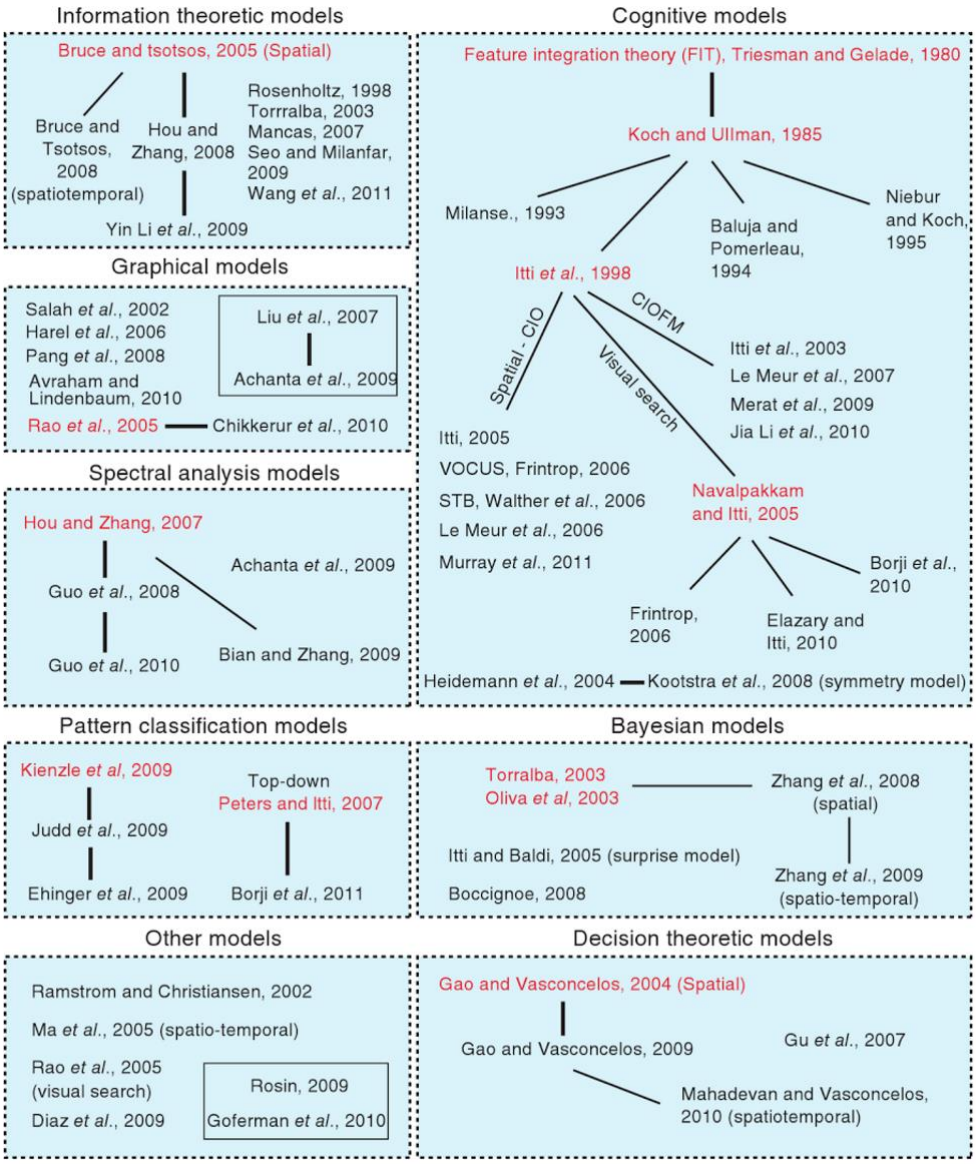


Figure 3-1: Taxonomy of the visual attention models. (Borji – Itti, 2013)

We focus our analytical part of the work on some specific parts of the visual attention modelling theory that formulate either the known global truth applicable on every approach to the research of visual attention or are good to understand to consider better the approach of our research. Among further discussed topics in following chapters we can find:

- spatio-temporal visual attention modelling,

- feature integration theory,
- cognitive models,
- learning-based models.

These are more in detail overviewed in the following subchapters.

3.1 Feature integration theory

History of the modern visual attention modelling is dated from the early 1980's when the first theory on the visual attention modelling was formed by Treisman and Gelade (Treisman – Gelade, 1980). In their work, they break down the visual attention modelling into the set of features which are affecting visual perception and forming our selection mechanisms. In the next step, they propose a method for combination of these features into the visual attention model. Continuum of this work was the research of Koch and Ullman (Koch – Ullman, 1985) which led to a novel method for combining these features and to introduction of the saliency maps. Considered features were elementary bottom-up features, like orientation of edges, colour, disparity and direction of movement, back that days. The scientists said, that conspicuity map of the scene, in the means of one of the features stated above, can be projected on new complex conspicuity map. When we repeat this process with every conspicuity map for every feature we are getting the saliency map. This process is illustrated on Figure 3-2. Definitions of the projections were not stated by the authors, yet.

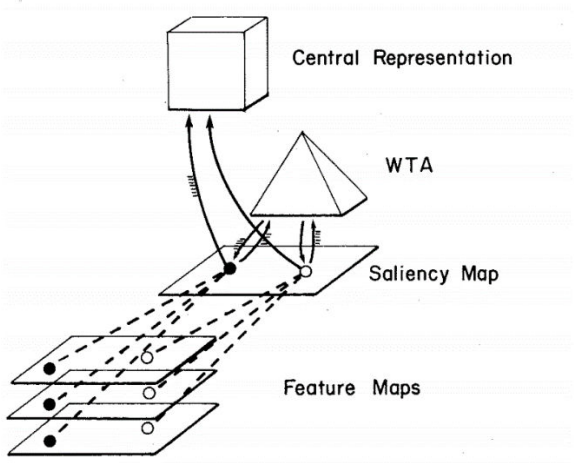


Figure 3-2: Process of combining the conspicuity (feature) maps into the single saliency map. Subsequently, the extraction of the most salient regions by the winner-takes-it-all algorithm is visualized. (Koch – Ullman, 1985)

3.2 Spatio-temporal visual attention modelling

Specific feature of visual stimuli is dynamics. Complex visual attention model should be able to deal with the features affecting visual attention in the spatio-temporal manner and the fact is that most of them does. There are several approaches known to include temporal information to a visual attention model according to Borji and Itti (Borji – Itti, 2013):

- Bottom-up models are most often enhanced with the ***motion*** information between two consecutive images of a captured scene.

- ***Flicker*** is an important characteristic temporal feature with an extensive research beneath. Flicker information often accompany the motion information from the captured scene.
- A few models build spatio-temporal models with respect to the information about the observation ***task and its progress*** through the time sequences of the captured scene.
- Temporal aspects can be extracted from the ***irregularities*** between consecutive scene images.

3.2.1 Motion

Motion between consecutive scene frames can be computed by the methods of computer vision, especially using the optical flow algorithms or spatially-shifted differences between Gabor pyramids.

Important base for working with the visual attention in egocentric video is the optical flow theory (or image velocity computation). Goal of the optical flow theory is computing an approximation of the two-dimensional motion field from the spatio-temporal patterns of the image intensity, assuming that the two-dimensional motion field is a projection of three-dimensional scene surfaces velocities on a two-dimensional camera surface (Barron et al., 1992; Verri – Poggio, 1987). There are different computational techniques for obtaining the optical flow algorithmically. Let us focus on the most widely used differential technique (Borji – Itti, 2013). The idea of the differential optical flow technique is based on the hypothesis saying that the intensity structures of time-varying image subsets are nearly constant for at least a short duration of time (time derivative). Let us define image intensity function depending on a time $I(x,t)$ according to the definition by Barron et al. (Barron et al., 1992). Then we can assume that:

$$I(x,t) \approx I(x + \Delta x, t + \Delta t)$$

where Δx is a displacement of image region I at (x,t) after time Δt. After applying Taylor series expansion:

$$I(x,t) = I(x,t) + \Delta I \cdot \Delta x + \Delta t I_t + O^2$$

Where ∇I is the special intensity gradient $\Delta I = I(I_x, I_y)$, I_t are the first-order partial derivates of $I(x,t)$ and O^2 represent second- and higher order of partial derivates which can be ignored. Finally, after a few mathematical operations and neglecting of O^2 we get the equation for the image velocity also known as the optical flow constraint equation:

$$\Delta I \cdot \boldsymbol{v} + I_t = 0$$

where $\Delta I = I(I_x, I_y)$ and $\boldsymbol{v} = (u, v)$ is the image velocity (Barron et al., 1992).

Differential method models cover two basic approaches to optical flow constraint- global methods and local methods.

Global methods interconnect the optical flow constraint with a regularization term (usually a smoothness constraint). Together, they produce a function which is minimized over an image region. When thinking of a regularization term in the means of a variance of the optical flow field we can pose a hypothesis that neighbouring velocities should be nearly identical (minimized) when these two belong to the same object (Barron et al., 1992; Horn – Schunck, 1981). This constraint leads to the definition of the error functional over the region of interest *D* in the image introduced by Horn and Schunck (Horn – Schunck, 1981):

$$\int_D \left((\Delta I \cdot v + I_t)^2 + \lambda^2 tr((\Delta v)^T (\Delta v)) \right) dx$$

Local methods make the use of the assumption that individual motion patterns are common. There was introduced the local constant model for the image velocity by Lucas and Kanade (Lucas – Kanade, 1981). They defined the image velocity as a weighted least square solution to the optical flow constraint. The velocity estimation is computed when minimizing:

$$\sum_{x \in N} W^2(x)(\Delta I(x,t) \cdot v + I_t(x,t))^2$$

where $W^2(x)$ is a square root of window function applied on an image and *N* denotes neighbourhood of a window function (Barron et al., 1992; Lucas – Kanade, 1981).

These methods are now often combined to provide better results as shown in the paper of Bruhn et al. (Bruhn et al., 2005). Authors prove in the paper, that local methods are more resistant to noise while, on the other hand, global methods are good at finding the dense flow fields. They combine the method of Horn and Schunck (Horn – Schunck, 1981) together with the one of Lucas and Kanade (Lucas – Kanade, 1981) into an optical flow model which yields the dense flow fields and is robust enough to deal with the noise at the same time.

3.2.2 Flicker

Flicker refers to continual presence and absence of the same stimuli at some frequency *f* on the scene and formally is defined as a difference of luminance intensities in the consecutive frames. Flicker belongs to low-level features affecting the visual attention and is often used to extract so-called flicker pyramids of the image (Itti et al., 2003).

Flicker paradigm is a phenomenon connected with change blindness research in neuropsychology. Change blindness means disability to detect changes in the scene (Goldstein, 2010; Rensink, 2002). Principle of the change blindness was discovered during an experiment in which identical images with one difference were presented to the observers with rapid alternations of a blank image. A series of the presentation periods were needed for the observer to notice the change between images. Until that time, the observer believed that the two images are the same ones.

Flicker fusion theory is based on the flicker paradigm using an assumption that changes in an image cannot be captured by the observer until the critical brightness versus the time-of-exposure combination is reached. This relationship is known as the Broca-Sulzer effect (Spillman – Werner, 2012) shown on Figure 3-3. The flicker fusion theory is applied for example in displaying devices. The devices change the displayed content in frequency higher than the critical flicker perception level. Flicker and its influence on the visual attention (as described in the previous paragraph) has to be taken into account of the visual attention modelling.

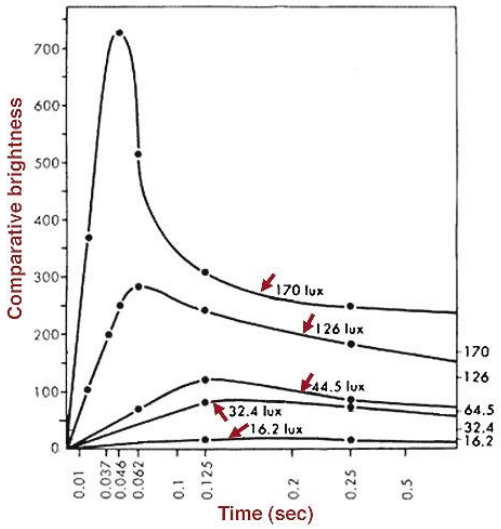


Figure 3-3: Comparative brightness of flickers with various light intensities in relationship with the time-of-exposure.[1]

3.3 Cognitive models

Cognitive models have its base in psychological and neuroscientific approaches to study human visual attention. Most of the visual attention models are inspired by the cognitive models or at least implement parts of them in some way (Borji – Itti, 2013).

One of the first very basic cognitive models which became the standard for model comparison is the evergreen work described in the paper of Itti et al. (Itti et al., 1998). The cognitive model proposed by Itti became very popular because it was proved by the years that it has high correlation with human real overt attention in the free viewing tasks. The model is based on the feature integration theory as it decomposes input image into a set of Gaussian pyramids and computes colour, intensity and orientation center-surround differences for each of the pixels in each of the pyramid stages. Colour center-surround difference is computed in respect to each colour channel of the RGB colour space. Colour, intensity, and orientation maps from each pyramid stage are, then, combined together resulting in the so-called conspicuity maps. The conspicuity map defines the saliency of the input image in the meaning of one distinct feature channel. These conspicuity maps are summed up in the end and normalized to obtain the single final saliency map for the input image (Borji – Itti, 2013; Itti et al., 1998). Visualization of the Itti's model proposal as described above is on Figure 3-4.

[1] KALLONIATIS, M., LUU, C. Temporal resolution. Available on 23/04/2018: http://webvision.med.utah.edu/book/part-viii-gabac-receptors/temporal-resolution/

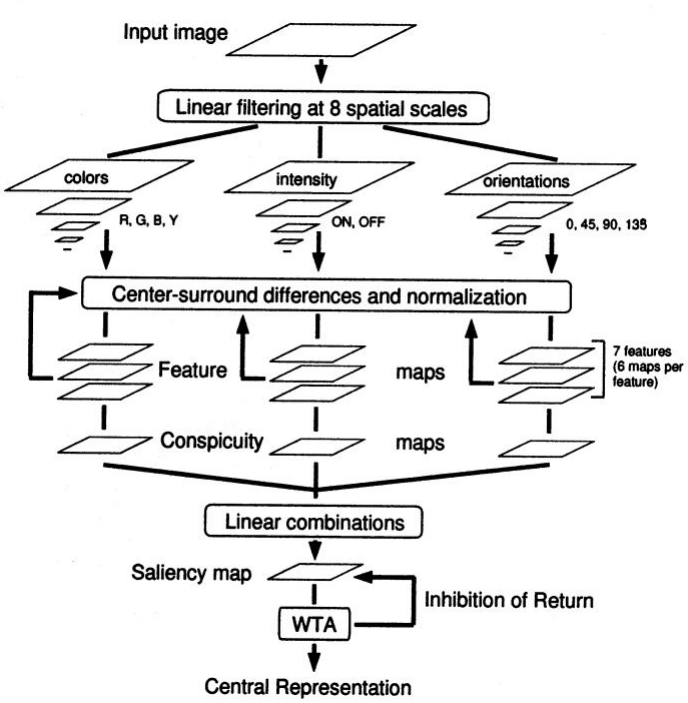


Figure 3-4: Itti's cognitive model of the human visual attention. (Itti et al., 1998)

Intensity feature maps of the Itti's model are computed by center-surround difference considering both two types of sensitiveness of mammal's neurons as:

$$I(c,s) = I(c) \ominus I(s)$$

where $c \in \{2,3,4\}$ and $s = c + \delta, \delta \in \{3,4\}$. Colour feature maps are computed from the RGB colour model representing "colour double-opponent" system typical for human colour perception. Neurons are excited by one colour and inhibited by another in the centre of their receptive field (Itti et al., 1998). Such colour pairs are red-green and blue-yellow. These colour double-opponent feature maps are computed as follows:

$$RG(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|$$

$$BY(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|$$

Orientation sensitive neurons in primary visual cortex can be represented by Gabor filters which are used as orientation representation in the Itti's model. Gabor filters are obtained from the Gabor pyramids $O(\sigma,\theta)$ in 9 scales ($\sigma \in [0..8]$) and under 4 angles ($\theta \in \{0°, 45°, 90°, 135°\}$):

$$O(c,s,\theta) = |O(c,\theta) \ominus (s,\theta)|$$

Obtained feature maps are combined into the mentioned conspicuity maps. Resulting saliency map is computed as linear combination of the normalized conspicuity maps $\bar{I}, \bar{C}, \bar{O}$:

$$S = \frac{1}{3}(N(\bar{I}) + N(\bar{C}) + N(\bar{O}))$$

Last step of the algorithm is the implementation of the winner-takes-it-all algorithm followed by the inhibition of return to the saliency map. This step identifies maxima intensity region in the saliency map which represents the most salient object. The inhibition of return removes the local maxima from the

saliency map (leaving it blackened) and allows us searching for the 2nd and following most salient regions on the image repeating the same algorithm (Itti et al., 1998). Ordering of the most salient regions on an image is useful, for example, to predict the amount of time needed for the observer to look at some specific object on the image (as we can see on Figure 3-5).

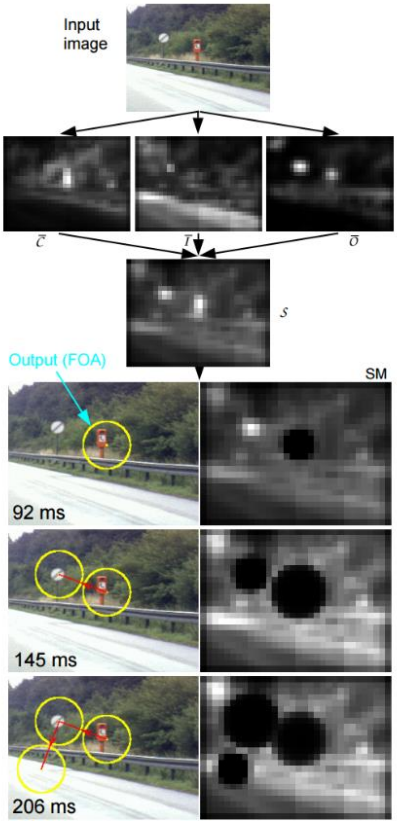


Figure 3-5: Itti's model of visual attention in an example. Saliency map is obtained from the conspicuity maps. Afterwards the most salient objects are extracted by the WTA algorithm and, finally, an approximation of fixation delays at the salient regions on the scene are computed. (Itti et al., 1998)

Cognitive methods were used during the research of more recent visual attention model by Le Meur et al. (Le Meur et al., 2007). Researchers set a goal to implement spatio-temporal cognitive model based on the same principles as the Itt's one (Itti et al., 1998). They defined set of achromatic, chromatic and temporal features which they extract from the input image and combine them together in the resulting saliency map. The asset of their work is mainly in including a temporal information in a cognition model. The architecture of the proposed model by Le Meur et al. is on Figure 3-6.

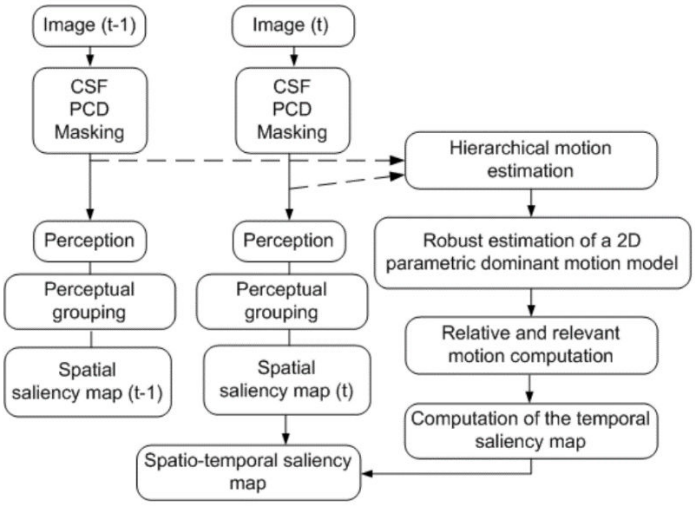


Figure 3-6: Cognitive model including temporal information from the motion. (Le Meur et al., 2007)

Team of researchers around Murray (Murray et al., 2011) opened a discussion about the integration of spatial features into cognitive models. They also point out the problem of justifying the choice of parameter values. Murray's team proposes a model which includes processing of visual stimuli according to the previous fixation pathways. They implemented inhibition mechanism, better reflecting the neural functions in the visual cortex, automatic adjustment of the center-surround inhibition window sizes, and specified parameters of the Gaussian Mixture Model on the eye-fixation data (Murray et al., 2011; Borji – Itti, 2013).

Cognitive models of the human visual attention help us not only in the modelling of visual system's biological behaviour, and to predict visual attention, but help us also during the further research of new facts about our neural mechanisms and its principles, namely object recognition (Borji – Itti, 2013).

3.4 Learning-based models

Machine learning can help us to build up a visual saliency model in different ways. It is most often combined with other types of the visual attention models to improve their performance or to enhance such a model with higher number of relevant features.

A Bayesian model, incorporating Bayesian machine learning algorithm, was introduced by Li et al. (Li et al., 2010). The proposed cognitive bottom-up model was enhanced with probabilistic Bayesian framework with an assumption that visual attention of people looking at the same scene is different and is defined by bottom-up features of the scene itself and the visual task. The saliency is in the probabilistic framework defined as:

$$p(E_{kn}|x_{kn}) = \sum_{t=1}^{T_k} p(T_{kt})p(E_{kn}|x_{kn}, T_{kt}) + p(T_{kb})p(E_{kn}|x_{kn}, T_{kb})$$

where E_{kn} is the event when the n-th macroblock of an image is the salient region in k-th scene in a video, $p(T_{kt})$ and $p(T_{kb})$ is the probability that the task T_{kt} or bottom-up feature T_{kb} controls the attention deployment in the k-th scene in the video and x_{kn} is a local descriptor representing the macroblock. To avoid difficulties in computing the probability $p(E_{kn}|x_{kn},T_{kt})$ we can substitute according to Li et al.:

$$y_{kn} = p(E_{kn}|x_{kn})$$

$$f_{kt}(x_{,kn}) = p(E_{kn}|x_{kn}, T_{kt})$$

$$b_{kn} = p(E_{kn}|x_{kn}, T_{kb})$$

and then we can make an approximation:

$$y_{kn} \approx \sum_{t=1}^{T_k} p(T_{kt}) f_{kt}(x_{kn}) + p(T_{kb}) b_{kn}.$$

Thus, we can obtain y_{kn} from the image, the eye-fixation data or labelled salient regions from an external source and train the machine learning model for saliency estimation (Li et al., 2010). Further discussion in the paper focuses on extending the saliency estimation function for multi-task events. For this purpose, the implementation of machine learning using kNN (k-nearest neighbours) algorithm is proposed. Visualization of the model design is on Figure 3-7.

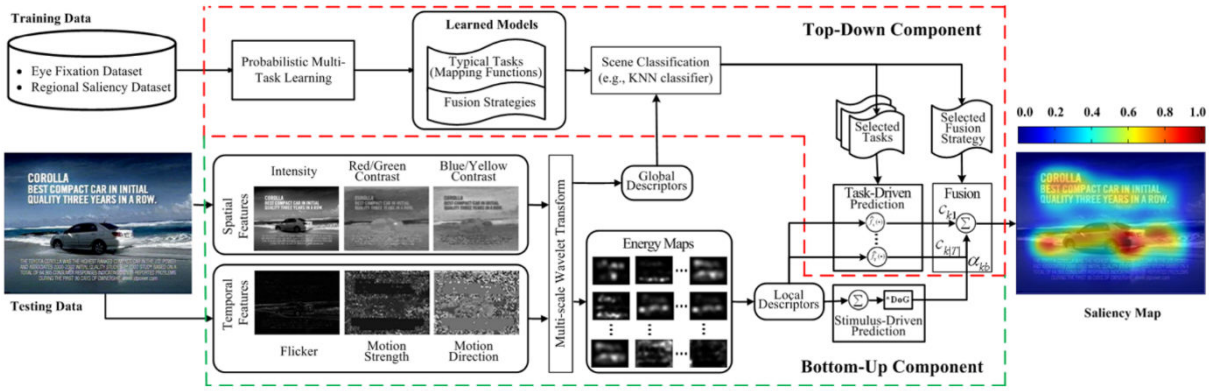


Figure 3-7: Visual attention model based on the machine learning. (Li et al., 2010)

Completely different approach in making the use of machine learning to predict visual attention is training the model directly from the fixation data. This approach was introduced by Kienzle et al. (Kienzle et al., 2009) whose work was based on the research of likeliness of the target on an image to become part of the saccadic movements in the free-viewing task. Team of researchers aimed to find a linear model fitting the eye-movement training data. They propose non-parametric bottom-up approach with non-linear mapping between an image patch and a fixation. The machine learning model expects positive output on the regions of fixations and negative output on random non-fixated regions on the scene (Borji – Itti, 2013; Kienzle et al., 2009). For each image patch of the size 13x13 pixels, they form 169-dimensional vector and fit the non-parametric model using the support vector machines (SVM) (Schölkopf – Smola, 2002) which is formally described by Schölkopf and Smola as:

$$f(x) = \sum_{i=1}^{m} \alpha_i e^{-\gamma||x_i - x||^2}$$

where x_i is the vector of characteristic features belonging to the image patch i, α_i is a weight for the image patch i and γ is a parameter to be set (optimal value was found out to be $\gamma = 1$). The machine

learning algorithm incorporates the extraction of the perceptive field which is according to Kienzle et al. (Kienzle et al., 2009) non-linear. They found out that the saccadic system can be interpreted by only four perceptive fields which can be seen on Figure 3-8. According to this conclusion, they defined the saliency model placing the radial basis functions centred at the perceptive fields:

$$s(x) = \sum_{i=1}^{4} \beta_i \varphi_i(x)$$

where φ_i is the radial basis unit $e^{(-\gamma||z_i - x||^2)}$ centered at the patterns z_1-z_4 from Figures 3-8 and β_i is a weight regarding the optimal values of γ (Kienzle et al., 2009).

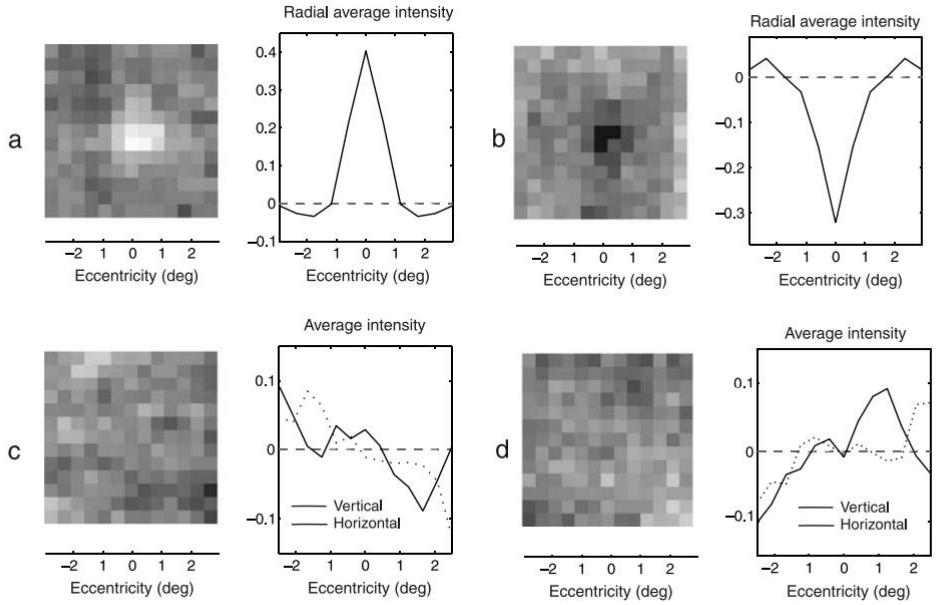


Figure 3-8: Non-linear perceptive fields which are the result of the machine learning algorithm appliaction in the visual attention modelling using support vector machines. Perceptive fields on (a) and (b) represent most-likely salient image structures and (c), (d) least salient image structures. (Kienzle et al., 2009)

4 State-of-the-art in visual attention modelling

There is an extensive ongoing research in the field of visual attention modelling regarding the past years. Many research papers introducing novelties in this field are presented on computer vision, psychological and neurophysiological conferences every year. Proof of this fact can be spotted on Figure 3-1 where different approaches to visual attention modelling are accompanied with the year of their introduction. The summary research paper summing up the state-of-the-art in the visual attention modelling is dated to 2013 which can be considered slightly out-of-date because the research in the field is becoming more and more extensive, nowadays. In this chapter, we try to reveal the current state-of-the-art in the visual attention modelling, overcoming the out-dated paper from the 2013. Mainly the novel approaches to visual attention modelling are inspiring us in choosing the proper research approach in this thesis. Therefore, to accomplish overview on visual attention modelling from previous chapter, we provide detailed overview on some of the most recent research.

One of the complete novelties is the application of the functional magnetic resonance imaging (fMRI) of human brains performing tasks related to the visual perception in the visual attention research. This approach is interesting to mention for demonstrating the importance of the modern technologies in the visual attention modelling. Reasons for the development of the model based on human brain's activity is related to the higher availability of the MRI devices, nowadays. This may be caused by the confirmed importance of the visual attention analysis in the diagnosis detection (Merker, 2007; Ungerleider et al., 2000). Speaking about the visual attention in the context of the diagnostics, Alzheimer disease diagnosis is under an extensive research and it was proved that the diagnose has strong correlation with changed visual attention patterns (Hao, 2005).

In the following subchapters, we discuss and analyse the most recent topics and novel approaches in the visual attention modelling:

- Including **depth information** (distance of the objects from the observer) into visual attention models and its correlation with neural responses to the three-dimensional scenes (Roberts et al., 2015).
- Visual attention research from the **egocentric perspective of the observer** (Matsuo et al., 2014; Buso et al., 2015).
- Visual attention and it's interconnection with **emotions** of people (Sripada et al., 2014; Hajcak et al., 2013).
- **Artificial intelligence and neural networks** as benchmark-leader's methods to predict the visual attention (Kümmerer et al., 2016; Kruthiventi et al., 2015; Vig et al., 2014).

4.1 Depth in the visual attention modelling

It is proved these days, that the third dimension of the scene (depth) plays a significant role as an aspect influencing our visual attention (Borji – Itti, 2013; Roberts et al., 2015). Crucial in the research in this field is to determine its exact impact on the perception of a visual scene. Subsequently, it is essential to involve depth in the modern visual attention models. The first research related to the depth influence on human visual attention was building upon the assumption that depth plays role in the pre-attentive stage of visual perception and, thus, it is essential to study its influence as the bottom-up aspect (Lang et al.,

2012; Wang et al., 2013). Nowadays, we know that depth is playing its role in the both pre-attentive stage and attentive stage of visual perception. Therefore, it is influencing visual attention in the bottom-up and top-down manners (Roberts et al., 2015). Influence on the visual attention in the bottom-up manner is truly straightforward but top-down effects of depth may not be so clear. We briefly introduce the reasoning of Roberts et al. to consider the depth influence in both manners in the next subchapter.

4.1.1 Depth as both bottom-up and top-down aspect

During visual searching tasks, users tend to segment the perceived scene and there is an assumption that understanding the process of segmentation can reveal us whether and how the depth of the scene influences top-down factors of the visual attention (Roberts et al., 2015). In the research work of Roberts et al., the two contradictory assumptions are further discussed: whether depth influences only the bottom-up manners of the visual attention, or it participates also as the top-down stimuli in the higher-order representation of the scene by segments. The results proved the second fact and Roberts et al. claim, that the depth's top-down influence on the visual attention may be even greater. The team of scientists held an experiment, in which numerous participants were asked to search for a particular letter of the English alphabet on an image with other letters. The letters were in different imaginary depth on the image while the target depth was either known or unknown. The depth illusion was achieved by tilting the two-dimensional image. Participants were provided with images containing 8 or 16 letters and the projection was either vertical or tilted (vertical for no-depth illusion) as we can see on Figure 4-1.

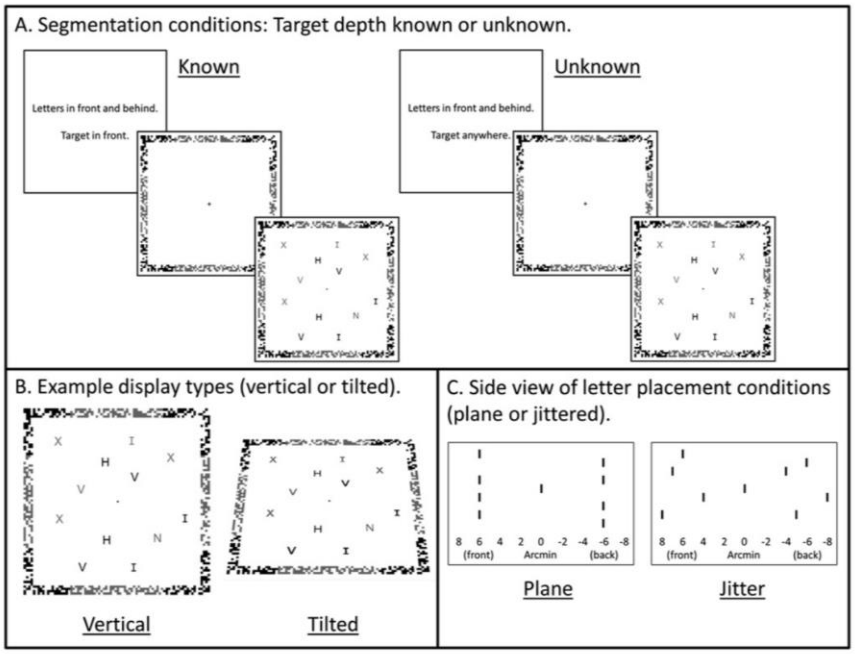


Figure 4-1: Examples of various image conditions presented to participants of experiment of (Roberts et al., 2015).

MRI data were collected during the user studies. Visual cortex responses of the specific cortical areas which should be sensitive to depth perception were monitored. Response times and accuracy results are visualized on Figure 4-2.

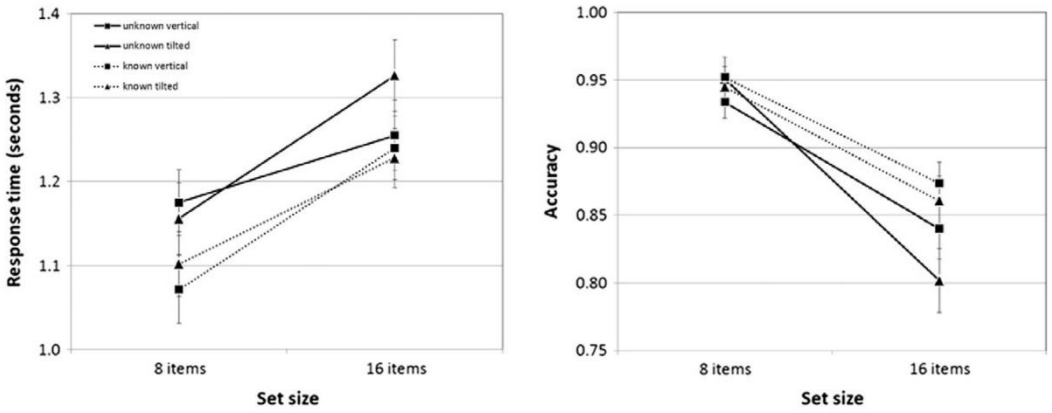


Figure 4-2: Response times and accuracy of response in an experiment with letter depths under different conditions shown on Figure 4-1. (Roberts et al., 2015)

We can see that people responded significantly more quickly when the information about the relative depth of the letter was known rather than unknown. Conclusion of Roberts et al. says that participants were, therefore, able to benefit from the relevant three-dimensional region (segment) of the image. Patterns of the activations in participant's brains are revealing us that target depth influences the activity in the depth-sensitive parietal regions, but not in the depth-sensitive visual regions. Based on this conclusion the assumption says that segmentation of the images with depth is the result of selectivity in the higher-order brain areas than in the perceptual regions (Roberts et al., 2015). Thus, according to Roberts et al., depth plays significant role both as the bottom-up and the top-down aspect and the top-down aspect may be even more significant than the bottom-up one. The top-down manner, however, can be effectively suppressed when conducting a user studies.

4.1.2 Depth influence on visual attention in real environments

Considering depth influence on the human visual attention, we partially build upon previous work of Olešová's master thesis (Olešová, 2016) where the research was focused on creating real world model of visual attention with included depth information. In this work an assumption, saying that depth is playing its significant role as the aspect of visual attention, was proved. A small unique dataset with a few participants was made throughout the user study involving free-viewing tasks on real scene in a laboratory.

Conclusion of Olešová states, that estimation of depth influence on visual attention can be approximal to polynomial function (Olešová, 2016). Thus, middle-depth location of objects in real world should take the most of human visual attention in neutral conditions:

$$2{,}5x^3 - 0{,}17 \cdot 10^{-3}x^2 + 0{,}03x - 0{,}75$$

where x is depth of the pixel in the image. However, the research is still open while the relevancy of the defined function representing depth influence on the visual attention was not proved, yet. We decided to continue in the research of the visual attention in a real-world-like conditions in a laboratory. However, based on the deeper analysis of the state-of-the-art in the visual attention modelling we propose novel and more innovative research approach.

4.1.3 Research of visual attention in three-dimensional scenes

There is a lack of datasets for studying depth influence on human visual attention in the real environments, or at least in three-dimensional scenes displayed to the observers. Two publicly available datasets of three-dimensional stereoscopic images displayed to the observers including depth information related to the image are described in this subchapter.

A computational model of stereoscopic 3D visual saliency – 3D Dataset. (Wang et al., 2013)

Work of (Wang et al., 2013) claimed in their paper that extensive and successful research in the field of human visual attention in the two-dimensional space (using common displaying devices and 2-D images projected to observers) is very popular and extensive till nowadays. They assume that the research should be extended to the three-dimensional space including information about depth of the scene provided to the observer. The research group around Wang was one of the initiators of such tendencies in the visual attention research. Back that days, the availability of datasets for the research of visual attention in three-dimensions was very poor. Wang et al. designed and created a novel 3-D dataset including the depth information. However, as they were one of the pioneers in the field, the dataset was not innovative and robust enough for extensive and more relevant research of depth influence on the visual attention.

The authors stayed with the static images displayed to the observers. The idea of the innovative approach was the fact, that the images were stereoscopic displayed through stereoscopic glasses. The created dataset contains two versions of each of the 18 static images involved in the user studies (one for the left eye and one for the right one) along with the disparity and depth maps, raw data from the eye-tracker and processed raw data into the fixation density map. Most of the images contain natural scenes, but despite of this fact, the dataset is not large enough and is not providing the observers real environments to study their visual attention in. Sample from the dataset is on Figure 4-3.

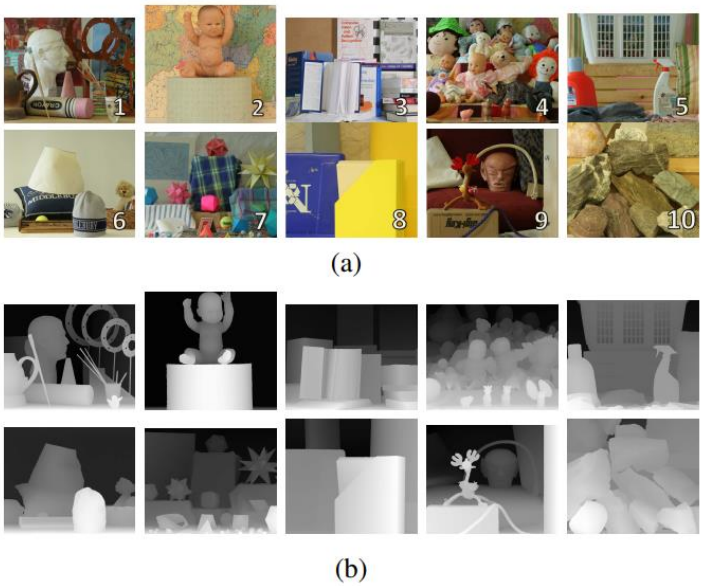


Figure 4-3: Sample of the 3-D dataset introduced by Wang et al. (Wang et al., 2013). Images in the group (a) were projected to the observers through the stereoscopic glasses. Corresponding depth matrices are visualized in image group (b).

Depth Matters – NUS3D-Saliency Dataset. (Lang et al., 2012)

There is another research work published in the field of modelling 3-D visual attention. The team of Lang et al. (Lang et al., 2012) did more significant step towards extensive studying of depth influence on visual attention by creating a novel, larger dataset to study human visual attention on.

The method proposal was very similar to the previous one- the dataset obtained during the user studies contains set of stereoscopic images (one for each eye) accompanied with depth maps and eye-fixation maps. The main asset of this dataset is the number of images (600) and number of participants (14) involved in the user studies. Researchers took static pictures of 600 real world scene images alongside with depth maps using Kinect device. From these two they computed stereoscopic images by proposed algorithm (more in Lang et al., 2012). They used common stereoscopic display throughout their user studies and projected the images to the participants- to some of them in 2-D and to some in 3-D. They analysed the correlation of the eye-fixation maps for 2-D and 3-D images during the evaluation phase. Sample from the dataset can be seen on Figure 4-4. The dataset is often used for comparison purposes of the novel visual attention models incorporating 3rd dimension of the scene. However, the dataset has some flaws. One of them are very noisy depth data which are sometimes not accurate and do not correspond with the image. Another flaw is that the dataset contains more images of the same scenes and that the relevancy of the ground-truth based on 14 participants divided into two groups is disputable. However, no better dataset including the depth information is publicly available, nowadays. Therefore, we will consider the dataset in the evaluation phase of our thesis.

Figure 4-4: Sample from the dataset created by Lang et al.(Lang et al., 2012). Top left: RGB image projected to the observers, top right: depth map, bottom left: fixation map corresponding to image projection in 2-D, bottom right: fixation map corresponding to image projection in 3-D.

4.2 Visual attention modelling and the egocentric video

User studies for the visual attention research were most often held in laboratories using common displaying devices with static eye-trackers monitoring gaze on the display. The things started to move forward along with introducing new technologies providing us new possibilities in the visual attention research. The main advancement in the research of visual attention is not only in incorporation of the

fMRI imaging but also in adopting egocentric video as starting point for new research possibilities. As fMRI helps us to understand how exactly visual perception works from the anatomical point of view, the egocentric video provides us the possibility to study the visual attention from more realistic point of view- view from the perspective of the human eye. This makes the research of visual attention in the real world (for example in people's everyday life) more possible and provides us the opportunity to move the research from the front of displaying devices towards reality. Egocentric video is an image sequence recorded "by an eye of observer" with a camera usually mounted on glasses of the observer capturing the scene influenced by the observer head's shakes and moves. Mobile eye-tracking uses the egocentric video principle and accompany the video with gaze information at each frame.

There are a few research papers incorporating the egocentric video and mobile eye-tracking. However, these are not focused on the research of human visual attention. Explicitly mentioned incorporation of the egocentric video in a research is mentioned in the paper of Matsuo et al. (Matsuo et al., 2015) where the team proposes a method for activity recognition in the egocentric video based on the attention. Another existing work tried to recognize activities in the egocentric videos by hand activities (Fathi et al., 2011) which are, however, insufficient in some cases because of the inability to recognize hands-free activities and activities where hands are not observed by the eyes. In the proposed algorithm of the state-of-the-art works, an egocentric attention map is built up in two main steps. Firstly, the gaze map is built as a grey-scale image in size of the original image where the value of each pixel is corresponding to a degree of the attendance at the pixel location. The next step is extracting observers motion and adjusting the gaze map to produce the final attention map. This method is briefly explained in (Yamada et al., 2011), which is main starting point in the visual attention modelling from the egocentric videos (Borji – Itti, 2013). Basis of this method may be described in an example in which an observer sees two equally salient objects on the left and on the right and his head moves to the right, so an object to the right should be considered more salient. Therefore, rotation and translation-based maps are defined and combined with gaze map as shown on Figure 4-5. We can build upon the knowledge of the equivalent saliency of objects on the scene and on the rotation and translation-based evaluation of object's saliency in the egocentric video in our research approach proposal in the next chapter.

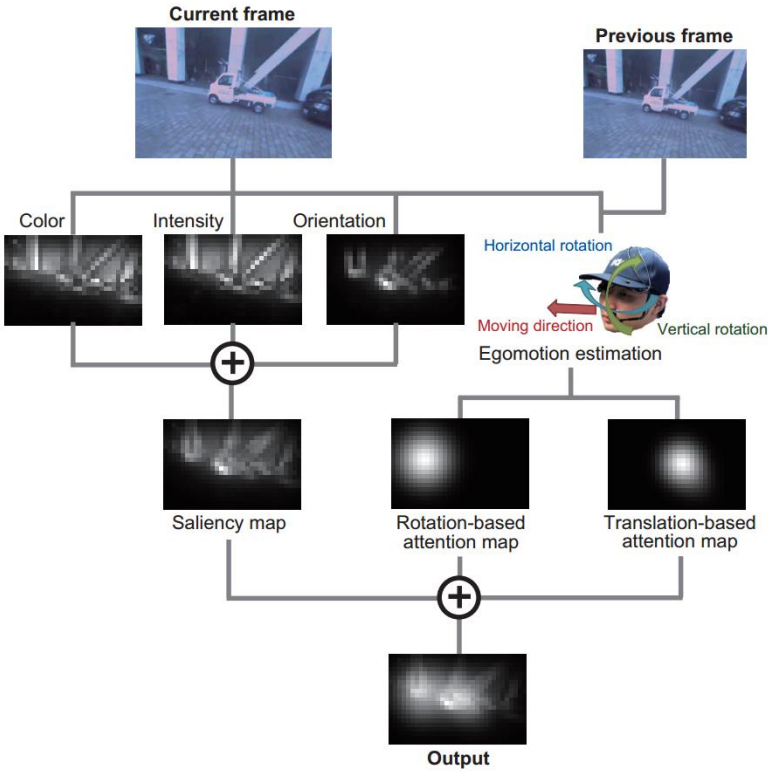


Figure 4-5: Design of the attention model based on the egocentric video proposed by Yamada et al. (Yamada et al., 2011) and used in the work of Matsuo et al. (Matsuo et al., 2015).

In the work of Buso et al. (Buso et al., 2015), the team builds upon the same starting point as Matsuo et al. and Fathi et al. as they focus on tasks in the egocentric video. Similarly, they consider only tasks in which an observer uses his hands. They neglect the recognition of the activity or task itself and focus on generating top-down saliency map based on the task an observer is working on with his hands. They define global and local descriptors of hands found in the egocentric video. Based on these descriptors, they define an object which an observer is interacting with. Subsequently, they assign the object some top-down saliency value by the proposed method. In their work they prove, that the computation of bottom-up saliency model was not useful at all during the tasks where hands were used. They claim that their method provides better results than current state-of-the-art bottom-up solutions. Evaluation was made only as a comparison to the existing state-of-the-art bottom up visual attention models as the approach in this paper is unique. The results are shown on Figure 4-6. For the future work, Buso et al. discuss combining their top-down task-driven model with some existing bottom-up saliency model including relevant temporal information.

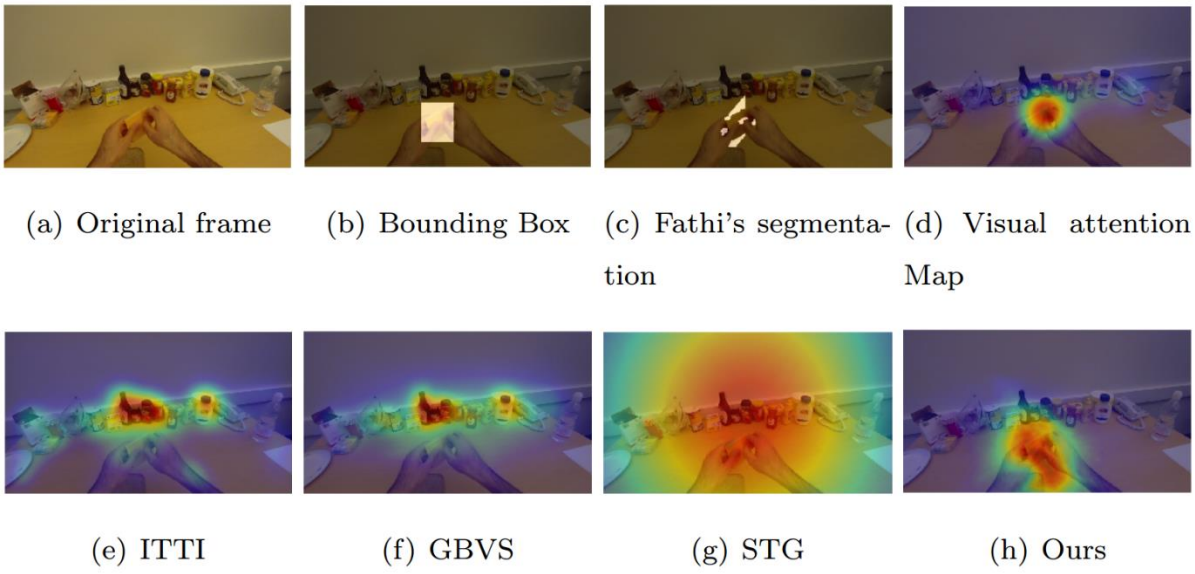


Figure 4-6: Comparison of the proposed model of Buso et al. (Buso et al., 2015) with the other bottom-up state-of-the-art models: (c) Fathi's segmentation (Fathi et al., 2011), (e) Itti's model (Itti, 1998), (f) Graph-based visual saliency model (Harel, 2007), (g) Spatio-temporal geometric model (Boujut, 2012).

4.3 Emotions and their effect on visual attention

Research, held during the past years, related to the emotions and their influence effect on the visual attention was focused on the emotions invoked by observing some image. It is proved that human brain is organized into many interconnectivity networks which have certain specific inter-relationships, and these alter during psychological tasks. One of them may be image perception and tasks related to it (Sripada et al., 2013). The Sripada's team conducts fMRI study on 54 participants monitoring specific brain's regions of interest and observing alternations in the interconnectivity networks while images with different Emotion Regulation Tasks (ERT) were presented to them. Emotion Regulation Tasks were validated during the previous research (Banks et al., 2007; Phan et al., 2005) and were related to images producing averse and negative emotions. The tasks were of different kinds, each related to emotion maintain and reappraise conditions. The maintain condition was the one when the participant was trying not to change the emotion invoked by the image and the reappraise condition was based on the instruction for the participant to knowingly change the emotion from the image to positive one. The fMRI results visualized on Figure 4-7 shows that the most significant changes in major of the interconnectivity networks in the brain (considering the difference between the maintain and the reappraisal conditions) are related to the visual network and dorsal attention parts with most changes, however, belonging to the visual network. Therefore, it is concluded that emotions and their changes are in strong relation with the human visual perception. Thus, the future study in the field of emotion sand their influence effect on visual attention is relevant.

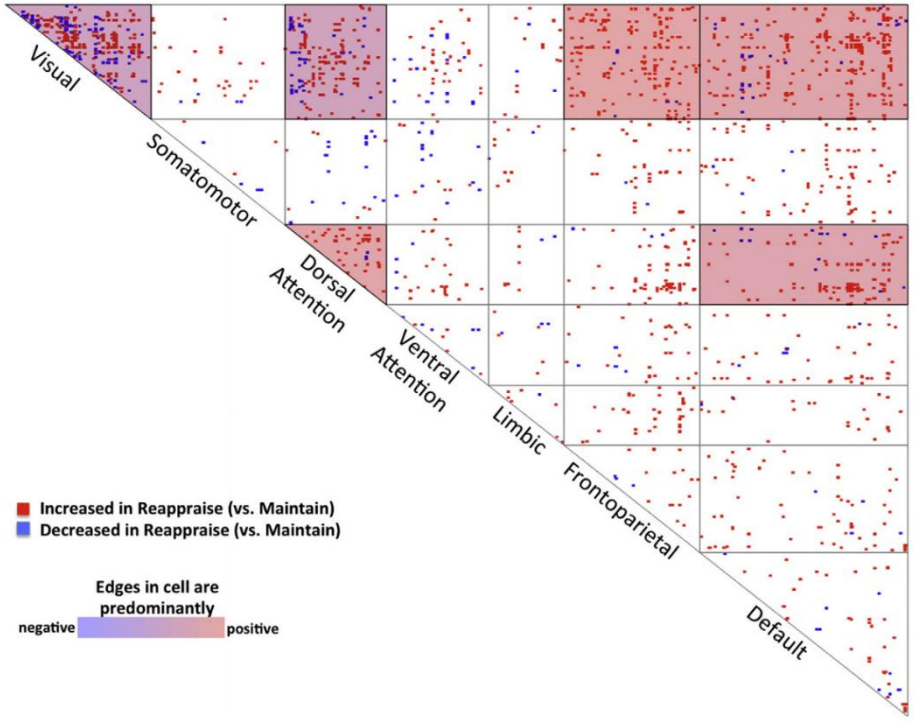


Figure 4-7: Cross-tabulation map visualizing changes in major of the interconnectivity networks in a human brain as a comparison between maintain and reappraisal conditions. (Sripada et al., 2013).

4.4 Artificial intelligence and neural networks in the visual attention modelling

A significant change in the visual attention modelling approaches is a shift towards the advanced learning-based computational models. This may be caused by the progress in biologically-inspired algorithms and methods which can now trustfully represent natural behaviour of the neural processes (Fister et al., 2013; Mirjalili, 2015; Mirjalili et al., 2016). The computational cost of the computer-vision methods applied on the high-resolution images is expensive on the resources. Therefore, learning-based modelling overcomes them in this point of view, too.

A fusion of the cognitive models with learning-based ones is clear from the research of Denil et al. (Denil et al., 2012) where an attention model driven by the gaze data is proposed. Neuroscientific theory of the visual perception is considered in the model in the means of two interactive pathways- identity ("what" aspect) and control ("where" aspect). Advantages of the neural networks with the images as the input were summed up in the research work of Mnih et al. (Mnih et al., 2014). They assume that the computational cost of the convolutional neural networks, applied on the high-resolution images, is too expensive on the resources. Thus, they propose a recurrent neural network able to predict the visual attention in the high-resolution videos with a satisfying time complexity.

Convolutional and deep neural networks can be applied in many areas of informatics and information science, among them in the field of the visual attention, too. First-of-a-kind fully convolutional network to predict visual attention was proposed by research collective of Kruthiventi (Kruthiventi et al., 2015). This convolutional network is able to learn multiple features needed to predict visual attention. The

learning is unsupervised, held in a hierarchical manner from training images in multiple scales. The global context is considered using network layers with large receptive fields. The network layers are proposed to be spatially invariant by implementing custom Location Based Convolutional Layer (LBCL). Architecture of the convolutional network is similar to the VGG-16 net (Simonyan – Zisserman, 2014). There are 7 blocks of which two are two-layered, three are three-layered (one is dedicated to high-level features and the rest are dedicated to low-level spatial features) and the last two are special blocks extracting semantic features from the images. Training parameters of the network are set according to Simonyan and Zisserman, initial learning rate of the first five blocks was set to $2 \, x \, 10^{-4}$ and learning rate of remaining blocks was set to $2 \, x \, 10^{-3}$ (Simonyan – Zisserman, 2014). The training phase is held in two phases. During the first one, the network is learning from the computed saliency maps in the SALICON dataset (Jiang et al., 2015). During the second phase, the convolutional network is learning from the actual eye-fixation maps (ground-truth) on different images than in the first stage. Kruthiventi with colleagues concludes that the main assets of their work are introduction of the inception module, extracting semantic features from the image, and the Location Based Convolutional Layer which is able to distinguish location-based patterns in the images. Combination of the mentioned assets can outperform the state-of-the-art models and highly correlates with the ground-truth of human eye-fixations as shown on Figure 4-8.

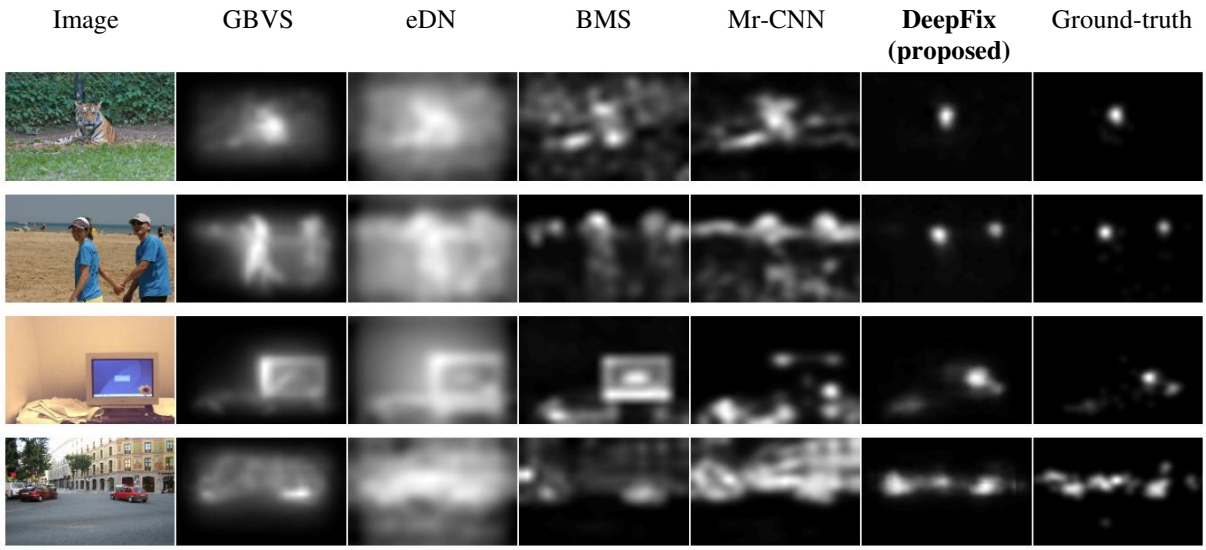


Figure 4-8: Comparison of the visual attention model using convolutional neural networks proposed by Kruthiventi (Kruthiventi et al., 2015) with other state-of-the-art models. The proposed model has significantly better results than the others and highly correlates with the ground-truth.

5 Proposed research approach

In this thesis, we introduce a novel method for conducting user studies focused on the research of human visual attention in real environments. We propose the novel method building up on the previous work and the visual attention modelling principles described in the analytical part of this thesis (Wang et al., 2013; Roberts et al., 2015; Olešová, 2016). The novel approach is based on suppression of the specifics of visual attention modelling from the camera perspective of view, which are well known from the research in the past years. In this thesis, the camera perspective is meant as the approach in the visual attention modelling where the images are displayed to the observers using standard displaying devices and the ground-truth data are collected as visual attention data referring to the image. We claim that the subsequent research on this kind of datasets then results in visual attention models predicting visual attention from the perspective of the camera.

Despite of the camera perspective, we propose to adapt a new perspective (the egocentric one) for the research of human visual attention. The ultimate goal of our novel research method is to implement our findings from the egocentric perspective of view into an existing saliency based on the perspective of the camera. The knowledge transfer between the modelling approaches is a great asset in the field of the visual attention modelling that we present in this thesis.

The novel approach leads to some novelties in the way we look at the saliency of the objects in the scene, too. The common bottom-up saliency model from the camera perspective looks at the static objects in an image through their planar properties as explained in the analytical part of the thesis, e.g. size, colour, orientation, contrast. Saliency of the object is computed by the saliency model based on these properties. However, taking into account real environments and the egocentric perspective of view (representing the perspective of a human eye), our approach looks at the same objects on the scene like on the objects with exactly the same properties influencing their saliency. This means that the objective bottom-up saliency of identical objects on the scene is equal. The subjective bottom-up saliency of the static objects perceived by the observer from his egocentric perspective is then greatly influenced by the object's position in the scene (and also by the internal state of the observer, a strong top-down aspect, which we will try to suppress in our research).

Based on the previous research in the field we claim, that the influence of two-dimensional position of the objects on the scene (meaning left-right position) on their saliency, as well as the center-surround difference influence, can be very accurately involved in the saliency computations by the state-of-the-art saliency models (e.g. the "DeepFix" described in the Subchapter 4.4 and more in Borji et al., 2013). Therefore, we will focus our research using the novel proposed method for conducting user studies mainly on the third dimension of the object positions in the scene and its impact on their saliency.

Moreover, we introduce the mentioned method for application of the novel findings from our research, following our novel method proposal, into the state-of-the-art saliency models based on the camera perspective. Speaking of this knowledge transfer, we claim that the saliency of the objects on the scene themselves (referred previously as the objective saliency) can be determined by the existing saliency models with high accuracy, too. Summing things up, the conventional saliency computations may lack only the information about the influence of the object's depth on the scene (the third dimension) on their saliency to predict visual attention closer to the visual attention in real environments from the egocentric perspective of view. Therefore, we focus the research following our proposed method on the depth influence on visual attention in real environments. Then, the aspect of the depth influence on the saliency

of the objects can be applied on the saliency map computed by the conventional saliency models. We propose this application in form of a weighting function- the depth saliency coefficient.

We study the depth phenomenon on a novel dataset created during the extensive user studies, following our novel proposed research method and regarding the novel research approach described in previous paragraph, and present the results of our novel approach and their evaluation in this thesis.

5.1 Research method overview

Goal of our proposed research method is to provide the possibility to study visual attention in real environments from the egocentric perspective of view. The method proposal is complete and detailed and contains description of all phases of the research described in the beginning of the Chapter- user study setup, methodology to conduct the user study, dataset creation and evaluation of the dataset.

Our proposed method is based on an idea to provide the observers, participating in the user study, a real scene in a laboratory and to collect data about the observer's visual attention during visual perception of the scene. This approach is in contrast with the previous methods for conducting user studies on human visual attention which provide an observer an artificial scene, generated by conventional displaying devices. Laboratory conditions were chosen to suppress the influence of surroundings in the exterior on the human visual attention as we want to focus on a specific bottom-up aspect influencing visual attention. Output from the user studies following our proposed method is a novel dataset for studying human visual attention in real environments including the information about the depth of the scene. This dataset is then used in our research of the depth influence on the human visual attention.

5.1.1 Real environments

Real environments are simulated in a laboratory which should be simplistic and with as less salient objects in the field of vision of the observer as possible. Regions of interest (ROIs)- in our method proposal 10 polystyrene balls hanging from the roof- are placed in different depths on the scene. Multiple standard LCD projectors are used to project various changing content onto the ROIs during the user study. Background of the scene should remain unaffected by the projection, so it does not distract the observer’s visual attention. Therefore, projection calibration procedure, using Kinect 2.0 device and methods of computer vision, is proposed to create a projection mask for each of the projectors. Details related to the calibration procedure are explained in the Subchapter 6.1.1. The experiment is proposed to be static, without motion influence on the visual attention. Therefore, an observer is required to stand still at a specific place during the user study. Proposed laboratory setup, representing the real scene, is illustrated on Figure 5-1.

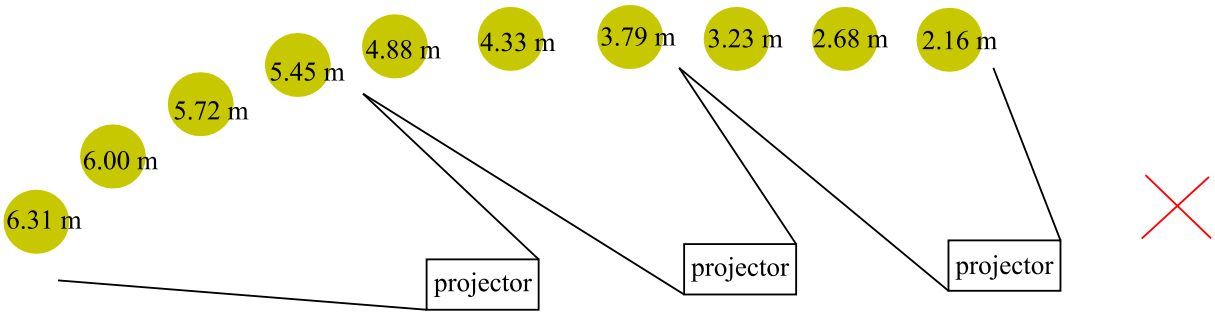


Figure 5-1: Schema of the user study setup in a laboratory. A participant stands at the red cross and yellow circles represents hanging polystyrene balls (our ROIs) with the distance from the observer written inside of them. We use three LCD projectors to cover all the ROIs by the projected content.

5.1.2 Capturing the egocentric data

Observer's visual attention data are collected during the user study by a mobile glass eye-tracker (more in the Subchapter 6.1.2). The observer is instructed to look freely on the scene while changes of the projected content on the scene are being handled by the projection handling software module (Subchapter 6.1.1). By free viewing task, we try to suppress certain top-down factors that influence the visual attention (visual searching, previous expectations, etc.). Hence, we want to observe the bottom-up influence of the depth on the visual attention. Caption from our final, targeted user study setup is displayed on Figure 5-2.

Figure 5-2: Participant, ready to start an experiment, captured along with the complete scene setup with desired projection on the ROIs. The eye tracker is mounted to a computer with running data acquisition module.

5.1.3 Projection overview

Content projected on the scene and changes of the projection may reflect goal of the research as the projection handling software module is proposed to be open for modification of the projection sequence. We propose our own projection sequence for the research of depth influence on human visual attention. The main specificity of our projection sequence is in projecting the same changing content on two ROIs on the scene at the same time (the action is further referenced as concurrent change on ROIs). Simplified, no content is projected on the ROIs at the beginning. Subsequently, the same content is projected on

two ROIs in different depths simultaneously, at the same time, for duration of 1000 milliseconds with 300 milliseconds fade-in and fade-out effect. There are generated complete $\binom{10}{2}$ combinations of these concurrent changes on the scene. The generated combinations have to be randomly sorted to avoid previous expectations during the user study. Based on the complete combinations, each ROI can be compared with each other during the dataset evaluation phase (e.g. in the means of the observer's first fixations immediately after changes on concurrent ROIs occur).

Three types of the concurrent changes on the scene and their random complete combinations are generated:

- projection of plain white colour,
- projection of slightly different colour tint,
- projection of the same face.

The projection sequence, it's visualizations and its development are explained more in detail in the Subchapter 6.2.3.

5.1.4 Proposed dataset overview

The novel dataset created during conducting the user studies, following our proposed method, should consist of:

- the egocentric video from the observer's perspective captured during the user study,
- the information about the observer's gaze referencing the egocentric video frames (in a separate file),
- the information about changes on the scene and their timing which can be synchronized with egocentric video.

Data and information that are valuable for further research and should be obtainable from the dataset by further data processing are, e.g.:

- fixations on ROI and their order after each change on the scene occurs,
- durations of fixations,
- delay of the fixations.

Details about the dataset are explicitly explained in the Subsection 6.1.2 along with the eye-tracking module proposal.

6 User study setup and methodology

We follow the previous Chapter 5 with detailed description of our proposed method and user study proposal from the technical point of view. The objective of the proposed method is to enable research of the human visual attention in real environments from the egocentric perspective of view. We already stated before that various specific hardware and methods of computer vision were used to fulfil our method proposal. All the used hardware and principles are further discussed in the next Subchapters.

6.1 User study setup proposal

The user study setup proposal follows the objectives of our research meeting the requirements of the novel research approach described in the Chapter 5. Goals of the user study setup are to:

- provide an observer a real scene in a laboratory as described in the Subsection 5.1.1 with projection aimed to study depth influence on the visual attention as discussed in the Subsection 5.1.3,
- monitor visual attention of the observer from the egocentric perspective of view as described in the Subsection 5.1.2,
- create a novel dataset as the output of the user study according to the dataset description in the Subsection 5.1.4.

Further requirements on the user study proposal to make the user studies repeatable, scalable and adjustable for further research of the visual attention in real environments are:

- automation of significant parts of the user studies,
- scalability of the user studies, maintaining the same laboratory conditions for each participant,
- adjustability of the projection sequence to match the needs of future user studies,
- supporting various content projection (colour, textures) onto the ROIs with various shapes.

We met all the goals and requirements in this thesis and included them in our user study setup proposal in the following Subchapters. Only easily accessible hardware in IT laboratories is used to conduct the user studies (with exception of the eye-tracker which is a piece of a specific hardware):

- common LCD projectors,
- Kinect 2.0 device,
- SMI mobile eye-tracking glasses.

The mobile eye-tracker is able to capture the egocentric video and gaze in the real-time. However, it is not able to capture depth information from the scene. No affordable, non-intrusive solution exists to capture gaze along with the egocentric video and the depth map of the scene at the same time during the user studies, nowadays. Therefore, we decided to make the whole experiment static (participant is standing the whole experiment at a specific place as captured on the schema on Figure 5.1.1) to avoid equipping participants with additional intrusive depth-capturing devices (e.g. Kinect). As the scene is static too, we can easily measure distances of the ROIs from the static participant for the evaluation purposes.

We propose projecting adjustable prepared content (projection sequence) on selected ROI on the scene to reach complexity of the scene and dynamic changes on the scene (see Subsection 5.1.3 and 6.2.3). The projection sequence will allow us great flexibility and mastery over the content appearing on the scene which can be targeted to study certain phenomenon of the visual attention. This is a great asset of the user study proposal as it can be reused in the future research of different aspects influencing visual attention.

There are numerous challenges related to reach goals of the user study setup described in the beginning of this subchapter. Among others, algorithms and methods of computer vision are used to meet them. The user study setup is, therefore, supported by three proposed software modules:

- projection management module,
- eye-tracking module,
- automatic evaluation module.

We describe the modules more in detail, along with the applied principles of computer vision, in the following subsections.

6.1.1 Projection handling module

Hanging polystyrene balls from the roof in different depths at the scene (meaning distance from a static observer) represent our regions of interest (ROIs) where the desired changing content is projected. We use binary projection mask applied on a conventional LCD projector's frame buffer with *1* values to project content on the desired ROIs and *0* values to project nothing on other areas at the scene covered by the LCD projector's projection plane. Thus, ROI background is unaffected by the projection and is not distracting the observer's visual attention.

We use multiple projectors across the laboratory to maintain good quality of projected content on each of the ROI at different depths. The desired state of projection on ROIs is captured on Figure 5-2. The computation of the binary projection mask (computation further referenced as projector calibration) is a challenging part of the setup which is further described.

Projectors are calibrated separately one-by-one using one Kinect 2.0 device[2] placed in the front of the device that is being calibrated. Working with multiple Kinect devices is not considered in the proposal as the Kinect device takes up more than 50% of the common computer's USB bus[3].

6.1.1.1 Projection calibration

Common LCD projectors are usually known as displaying devices, being able to project desired content on a homogenous white background. Not so well-known usage of the LCD projectors is to make them project directly on desired areas on scene (which are within scope of projection plane), transforming the projectors to multifunctional reflectors. Under projection on the desired areas we understand projection of black colour on non-ROI areas and any other colour on ROI areas. This creates the illusion of projecting content on specific areas at the scene using e.g. theatre reflectors. This illusion is working due to the fact that the common LCD projectors project black colour- *RGB(0,0,0)*- as "nothing". The illusion can be easily proved by visual experiment with an LCD projector which we made as a feasibility

[2] Manufacturers website: http://www.xbox.com/sk-SK/xbox-one/accessories/kinect
[3] Kinect 2.0 documentation available on 6/4/2018 at: https://msdn.microsoft.com/en-us/library/jj131023.aspx

study in the beginning of our research. Result of the test is captured on Figure 6-1. Projection calibration necessary for obtaining the projection mask applicable on the projector's frame buffer is described in the next paragraphs.

Figure 6-1 Testing the illusion of projecting bright light on specific areas at the scene (marked by red circles) using e.g. theatre reflectors. In fact, LCD projector projects white colour on the ROIs and black colour on non-ROI areas which creates the mentioned illusion.

Kinect device should be placed in the front of the LCD projector that is meant to be calibrated. The Kinect should be aligned with the projector, so it's camera lenses are in the same direction and angle as the projector's lens to simplify the calibration procedure. The information acquired through the Kinect device are the RGB frame and the normalized depth map of the scene in the range $\langle 0;1 \rangle$.

The RGB frame is used for segmentation of the projection plane by its corner coordinates using either human interaction (clicks on corners) or an improved adaptive Gaussian mixture model for the background subtraction (Zivkovic, 2004). Projection plane can be easily segmented from the RGB image when the Kinect is aligned with the projector (Figure 6-3). Using this simplification in the calibration process, we avoid more complicated geometric transformations between coordinate spaces and we may focus on the main objective of this thesis.

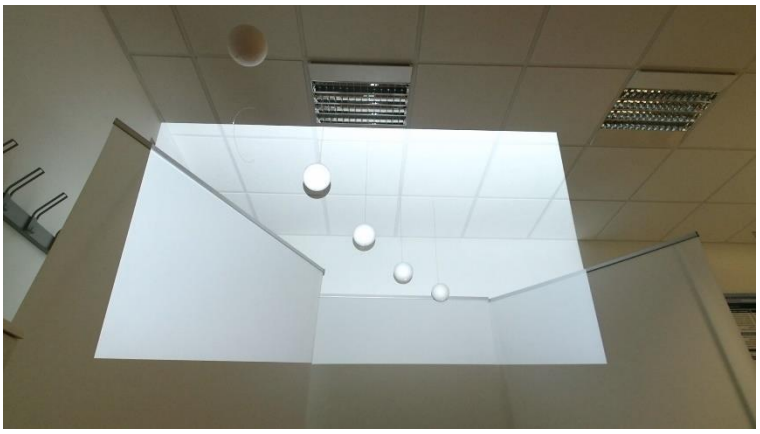


Figure 6-2 The RGB frame from the Kinect with the visible projection plane that can be easily segmented from the frame. Projection of plain white colour during the calibration is handled by the calibration module.

The Kinect's depth map space coordinates do not match the Kinect's RGB image space coordinates. Therefore, we have to define homography projection between the RGB space and the depth map space

and vice versa. This homography can be defined using depth point-cloud space of the Kinect as proposed by Khoshelham-Elberink. The depth point-cloud space can be used for the projective transformation of the projection plane in the RGB space to the depth map space and transformation of the ROI locations in the depth map to the RGB image space. We obtained planimetric object coordinates $P(X_k, Y_k)$ of each point in the image from its image coordinates $P(x_k, y_k)$, the scale determined by the distance (depth) of the point k in the object space Z_k and the focal length of the camera f:

$$P(X_k, Y_k) = \frac{-Z_k}{f(P(x_k, y_k) - P_0 + \delta_{P(x_k, y_k)})}$$

where P_0 is the principal point, and $\delta_{P(x_k, y_k)}$ are corrections for lens distortion (Khoshelham-Elberink, 2012). We can perform projective transformations between points in the RGB frame and points in the depth map using the defined homography. Thus, we can define the segmented projection plane in the depth map. Projective transformation of the whole depth map to the RGB image space by the inverse homography defined above is visualized on Figure 6-4.

Figure 6-3 Visualization of the homography transformation of the depth map (with ROI locations marked by red colour) to the RGB image space using the inverse homography matrix of the RGB frame projection to the depth point-cloud space. If you place the images one on another, the positions of the pixels would match and refer to the same point in the depth point-cloud space.

The ROIs at the scene are proposed to be of non-salient colour (the same as the background). Thus, ROI segmentation from the RGB frame would be difficult (see Figure 6-4 on the right). Therefore, projection plane segmented from the Kinect's depth map is used for segmentation of the ROIs from the background of the scene. The ROIs stand out in the depth map as they are located in different depths than the background is.

We incorporate human interaction (clicking on ROI areas) and flood-fill algorithm (flood-filling clicked depth level in the depth map) in the segmentation process (Figure 6-2). The user interaction during the calibration process is appropriate while the ROIs can be of different shapes and in different depths during the future user studies. We try to maintain our setup proposal general enough for its future use in this case. Only the segmented-out projection plane from the depth map is displayed to the user during the ROI segmentation for better usability of the user interface. It is obvious that only the ROIs within the projection plane may be affected by the projection of the LCD projector.

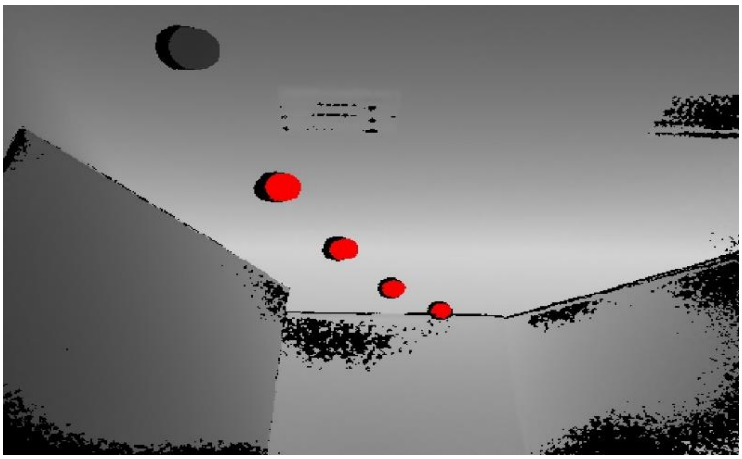


Figure 6-4 Visualized projection plane part of the Kinect's depth map with segmented ROI locations by incorporating the user interaction (clicks on the ROI areas) and the flood-fill algorithm.

Binary projection mask *M* in the depth map space is obtained after the ROI segmentation using flood-fill algorithm as:

$$M(x,y) = \begin{cases} 0, & if\ \ I'(x,y) - I(x,y) = 0 \\ 1, & otherwise \end{cases}$$

where I' is the depth map with the flood-filled regions and *I* is the original depth map before the flood-fill operations. Projection mask in the depth map space is transformed back to the RGB space by the homography using the depth point-cloud space.

Projection space of the LCD projector (or projector's frame buffer space) is relative to the Kinect's RGB space. Thus, we can compute another homography between the Kinect's RGB space and the projection space. This is an orthogonal projection between two two-dimensional spaces. Therefore, we can use the projection plane corner coordinates from the RGB image to define a set of four equations with one unknown homography projection matrix *H*:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = H . \begin{pmatrix} x \\ y \end{pmatrix}$$

where the projection plane corner coordinates from the RGB image are substituted for *x* and *y* and $x', y' \in \{[0,0], [0,1], [1,0], [1,1]\}$. The homogenous homography matrix between the RGB image space and the projection space can be determined by solving the set of equations using principles of the algebra.

Homography transformation using the matrix *H* applied on the projection mask in the RGB image space results in the projection mask in projection space of the LCD projector (projector's frame buffer space). The final, double transformed projection mask can be applied on the projectors frame buffer to project content directly on ROIs (Figure 6-5). Whole calibration process is repeated with all the LCD projectors. User interaction with the projection handling module is modelled in the sequence diagram on Figure 6-6 (considering the interaction during the projection plane segmentation, too).

Figure 6-5 One calibrated LCD projector with projection mask applied on its frame buffer is projecting on the desired ROIs.

There is a re-calibration step involved in the calibration process after obtaining the projection mask applicable on the projector's frame buffer. During the re-calibration step the RGB frame of projection on the ROIs is acquired by the Kinect device (as on Figure 6-5). The module automatically computes the projection error from the calibration phase, comparing the captured frame with the ROI locations segmented from the depth map space and transformed to the RGB space earlier. ROI locations are segmented from the captured RGB frame as regions with the highest pixel intensities, having volume which may approximately correspond to the ROI in the frame. The detected ROIs are then labelled from the left-hand-side, starting from the zero, to match the labelling of the ROIs during the calibration. The re-calibration procedure matches ROIs segmented from the captured RGB frame with the ROI locations segmented from the depth map space and transformed to the RGB space earlier. The projection error is then computed for each of the ROIs separately. The error denotes the shift vector applied on the ROI coordinates in the resulting projection mask.

If one is not satisfied with the result of the projection after the calibration phase (including the re-calibration step), the projection mask can be manually fine-tuned. There is a user friendly interactive interface for the fine-tuning of the projection with immediate responses on the scene when changes in the projection mask occur (for more details see Appendix B of this thesis).

Calibration data for each projector are saved on the disk in separate binary files containing the calibration data structures. Saved binary file can be read by the projection handling module. This allows us to use only one Kinect device to calibrate all the projectors one-by-one, to transfer the projection data to another computer with projection handling module, or to reuse the calibration next time when needed (if the object placement in the laboratory does not change). If the laboratory setup shifted a little bit from the time of the last calibration, the proposed manual fine-tuning of the projection may be useful after loading the saved calibration data, too.

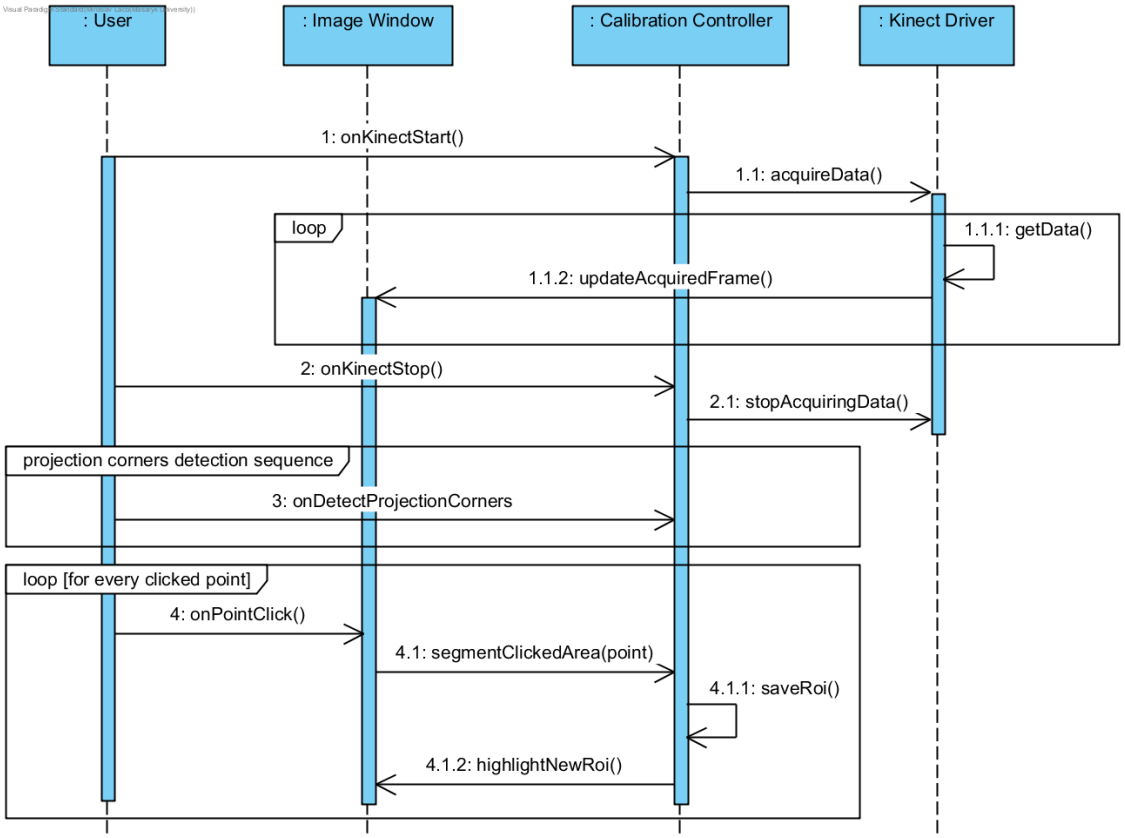


Figure 6-6: Sequence diagram of all the required user interaction with projection handling module throughout the calibration process.

6.1.1.2 Content projection

We have the information about ROIs position in the depth map space, RGB frame space and the frame buffer's projection space. It is trivial to detect the ROIs in the projection mask and label them from the left-hand-side to the right-hand-side beginning from the zero. The labelling will hold for the ROI positions in all the other projection spaces, too. Displayed ROI labels are visualized on Figure 6-7.

Figure 6-7: Labeled ROIs segmented from the projection mask and transformed to the RGB frame space for visualization purposes.

Similarly, labelling the projectors used in the user studies would result in uniform differentiation among them. Hence, the ROI on the scene can be unambiguously referenced by the projector label combined with the ROI label. Each ROI should be for practical reasons referenced by the unique number starting

from the zero to make the projection sequence easy to define. For this purpose, the hash map is defined. It translates the unique ROI labels onto the projector-ROI label combinations. The unique ROI labels are assigned to ROIs by the convention starting from the zero from the left-hand-side.

The application of the projection mask on the projector's frame buffer is handled by the operations on graphic chip of the computer using OpenGL. Each labelled ROI in the projection mask is treated as an OpenGL polygon. Thus, any colour or texture may be interpolated within the polygon instead of projecting plain white colour on ROI through the projection mask. Finally, projection module is prepared for the projection sequence rendering, as we can reference any of the defined ROIs on the scene and interpolate any content within it's coordinates in the projection mask.

Projection sequence handled by the module is adjustable. It is possible to define a unique own projection sequence in the form of ROI labels with corresponding content that should be projected on them accompanied with hold duration of the change. There can be any number of such changes contained in one projection sequence. Defined changes are all rendered by the module and the transitions between them are generated automatically so the changes on the scene are not too sudden. By default, the fade-out and fade-in effect (each with duration of 300 milliseconds) is added between the two changes. The rendered sequence is then projected on the ROIs whole at once, using the module's "play" directive. New projection can be started right after the previous one is finished. More about the usage of the module and the projection sequences can be found in the Appendix B of this thesis.

6.1.2 Eye-tracking module

We used the mobile SMI Eye Tracking Glasses[4] (equipped by the observer on Figure 6-8) incorporating three-point calibration procedure[5] provided with the SMI SDK for Windows platform[6] to acquire the egocentric frames at *60Hz* sampling frequency with the corresponding gaze data during the user studies. There is a free software application iView ETG provided by the SMI company, running on Windows platform, compatible with these glasses. The software provides driver for the eye-tracker and simple user interface with only limited possibilities of customization of the data gathering process. The main shortfall of the software is the impossibility to store the egocentric video and the gaze information in two separate files. The gaze is hard-written in the acquired egocentric video frames. This type of output is good for visual analysis and evaluation of the dataset. However, it is not suitable for further processing of the gaze data or automated dataset evaluation. Therefore, the eye-tracking module based on the SMI SDK is proposed for the data acquisition part of the user study.

[4] Manufacturers website available on 28/04/2018 at: https://www.smivision.com/eye-tracking/product/eye-tracking-glasses/
[5] Explanation of 3-point calibration of the eye-tracker available on 24/02/2018 at: \url{http://tsgdoc.socsci.ru.nl/images/c/cb/IView_X_SDK_Manual.pdf
[6] Available on 24/02/2018 at: \url{https://www.smivision.com/wp-content/uploads/2016/10/smi_prod_sdk.pdf

Figure 6-8: The observer equipped with the mobile SMI Eye Tracking Glasses. The eye-tracker is connected through the cable to a computer with the running eye-tracking module.

Architecture and implementation of the SDK is noticeably outdated and does not contain many of the features implemented in the SMI software products. The SMI company, however, does not exist anymore and does not provide any software purchases or SDK updates since 2017. Therefore, we propose our own eye-tracking software module built upon the existing SMI SDK. It has client-server architecture where:

- server-side is acquiring raw data from the eye tracker and handling the data requests from the client,
- client-side is sending data requests to the server and processing the received data.

The requirements on the eye-tracking module are:

- writing the egocentric video frames into the video file with common encoding, and with frame-rate equal to the sampling frequency of the eye-tracker (*60Hz* corresponds to *60 frames-per-second* in this case),
- writing gaze into a separate structured file so the gaze samples and the egocentric video frames can be easily matched.

Our module, extending the client-side of the SDK, is able to save the egocentric frames from the eye tracker into a video file with *wmv* format. We chose this video format as it is as loss-less as possible and, at the same time, it is widely supported across the Windows platforms. Raw gaze data are saved by the client-side for each video frame in a separate structured *csv* data file. The egocentric frames saved in the video are referenced by each gaze sample in the data file. This way, further data processing of the gaze is possible without modification of the egocentric video. The activity diagram of our proposed extension is on Figure 6-9.

As mentioned in the previous paragraphs, gaze data from the eye-tracker are raw and lack any post-processing on the SDK's side. However, we decided to store them in the data file as raw samples, so we

do not drop any frames resulting in the frame-rate drop in the video. We proposed and implemented a post-processing algorithm to filter out noise and to deal with invalid data samples in the gaze data. The post-processing takes place after the user study is finished and the gaze data from one participant are complete. The input for the post-processing algorithm is only the data file with gaze information.

As the algorithm, we adapted and implemented part of the I-VT fixation classifier algorithm published by Olsen (Olsen, 2012). We chose only suitable parts of the algorithm for our post-processing procedure as the paper is only partially relevant for the mobile glass eye trackers and originally was meant for the static eye-trackers under the display. The relevant proposed and implemented algorithm parts are:

- gap fill-in algorithm using linear interpolation between valid samples,
- noise reduction algorithm using median filter with small window size (3 gaze samples) as proposed by Olsen.

The post-processed gaze data file is ready to be an input for further data processing, visualization of the egocentric video with written gaze (by our simple proposed software tool), and for the automatic evaluation module.

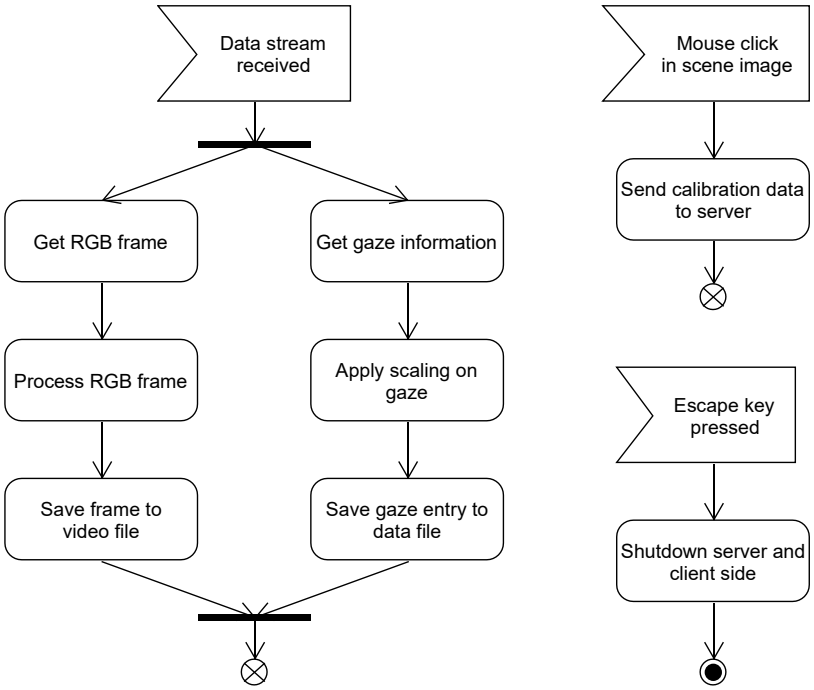


Figure 6-9: Activity diagram of the event handling on the SMI SDK extension's client-side proposed in this thesis.

6.1.3 Automatic evaluation module

The automatic evaluation software module is proposed to make the evaluation of the information, obtainable from the dataset by further processing, simpler and faster. The information possible to mine from the dataset are: fixations on ROI after change of the scene, their order, duration, delays, etc. We focus our research on the bottom-up aspect of visual attractivity of the changes on the scene captured in the egocentric video. As we know from the visual attention theory (Chapter 2), the bottom-up aspects

affecting visual attention are the strongest during observing the scene (or a change on it) for the first milliseconds. Therefore, we found fixations on ROIs after change on the scene and the fixation's order as the most valuable information for our research among all the obtainable information from the dataset. Thus, we can easily focus on the first fixations after change on the scene in the evaluation phase. The goal of the automatic evaluation module is, therefore, to provide an information about fixations on ROI for each frame of the egocentric video in the dataset.

The input for the automatic evaluation module are the data acquired by the eye-tracking module and post-processed using our post-processing algorithm (described in the previous Subchapter):

- video file containing the egocentric video frames,
- corresponding data file with post-processed gaze information for each video frame,
- projection sequence protocol.

The only information we lack from the dataset to get our desired output in a straightforward way are locations of our ROIs (their border coordinates) at each egocentric video frame. We propose two methods for ROI segmentation from the egocentric video:

- segmentation of the image regions with the highest pixel intensities,
- enhanced segmentation using fiducial markers (Garrido-Jurado et al., 2014).

6.1.3.1 Segmentation of the image regions with the highest pixel intensities

Segmentation of the image regions with the highest pixel intensities is a simple and fast segmentation algorithm. The algorithm has an assumption that bright light (white colour) is projected on all ROIs at a frame on which the segmentation takes place (e.g. white colour projected as a calibration before starting the projection sequence). The enlightened ROI should then be the brightest objects on the scene. We can afford such an assumption because of low brightness conditions in laboratory throughout the experiments (for more details see Subchapter 6.2.1). The segmentation algorithm has two steps:

- image thresholding with high constant threshold value, computed as a maximum pixel intensity in a frame from which is subtracted a constant tolerance value (can be modified manually),
- contour detection algorithm introduced by Suzuki (Suzuki et al., 1985), applied on the output of the threshold operation, where detected contours represent ROI border coordinates in the egocentric video.

6.1.3.2 Enhanced segmentation using fiducial markers

Fiducial markers segmentation is used in computer vision to precisely detect coordinates of the marker position in an image (Garrido-Jurado et al., 2014). Therefore, we assume that our segmentation algorithm should be more reliable and robust using them. Fiducial markers are projected on the ROIs during the eye-tracker calibration step and segmented from the egocentric video frame instead of using simple segmentation algorithm described in the previous Subchapter. The fiducial marker segmentation algorithm implementation by Garrido-Jurado is publicly available. After obtaining the accurate ROI coordinates in the frame, the enhanced segmentation step should be followed by the previously proposed segmentation of the image regions with the highest intensities to determine border coordinates of the ROIs.

The enhanced segmentation algorithm is proposed and implemented as part of this thesis. However, it was not used during our user studies because this method is not indifferent on marker deformations

which do not preserve straight lines- and the deformation of projection on polystyrene balls does not. Therefore, marker content projected on the polystyrene balls was undetectable. The enhanced segmentation may be used for the future user studies where flat ROIs are used.

6.1.3.3 ROI tracking in the egocentric video

We decided to track the ROI locations between two consecutive frames of the egocentric video instead of running the time-consuming segmentation algorithm frame-by-frame. We experimentally studied two types of flow algorithms (global and dense optical flow) used for tracking objects between video frames in computer vision applications. By observations of results in relation to high specificity of egocentric video (head shakes, tilts and moves, depth of the scene, etc.), we found out that global optical flow is not matching our needs. Therefore, we suggest to involve dense optical flow algorithm by (Farnebäck, 2003) in our ROI tracking algorithm proposal. The algorithm computes flow (a shift vector) for every pixel of two consecutive frames. The algorithm works well on the edges. However, it does not work well on non-edge homogenous regions where the flow cannot be determined. Visualization of the flow applied on the egocentric frame is on Figure 6-10.

Figure 6-10: Visualization of the dense optical flow in the egocentric frame. The green arrows denote vectors of flow at specific pixels. Notice, that flow is computed correctly only around edges of objects.

We compute the approximal shift vector for each ROI separately using the flow map. The computed shift vector is added to the border coordinates of the ROI from the previous frame to obtain the new ROI coordinates in the consecutive frame:

$$C'(x', y') = C(x, y) + \vec{v}_{shift}$$

where C is a set of the ROI border coordinates in current frame, C' is a set of ROI border coordinates in the succeeding frame and $\vec{v}_{shift}$ is the computed approximal shift vector of the ROI. The shift vector $\vec{v}_{shift}$ is obtained as the mean flow vector of the pixels around the ROI border coordinates:

$$\vec{v}_{shift} = \overline{(F \cap C_{surr})}$$

where C_{surr} is the union of C with the pixel coordinates around the ROI border coordinates (border surroundings), in an absolute distance of δ pixels from C in the current frame, and F is the flow map

between current and consecutive frames calculated by the dense flow algorithm proposed by Farnebäck. We consider C_{surr} instead of just the C set in the computation of the ROI shift vector due to better results when the flow map is noisy directly on the ROI border coordinates. We also cannot consider C with all the pixels inside of the ROI due to distorted values of dense optical flow inside the ROI with homogeneous content.

The shift vector approximation cumulates an error during the ROI tracking phase. We found this issue during the prototyping phase of the proposal (see Figure 6-13). The error is eliminated every n video frames by repeating the segmentation algorithm on the n-th frame. There are more precise methods for object tracking in the egocentric video- for reference see (Ren et al., 2010). However, they are not publicly available, nowadays, and their proposal is an untrivial research task.

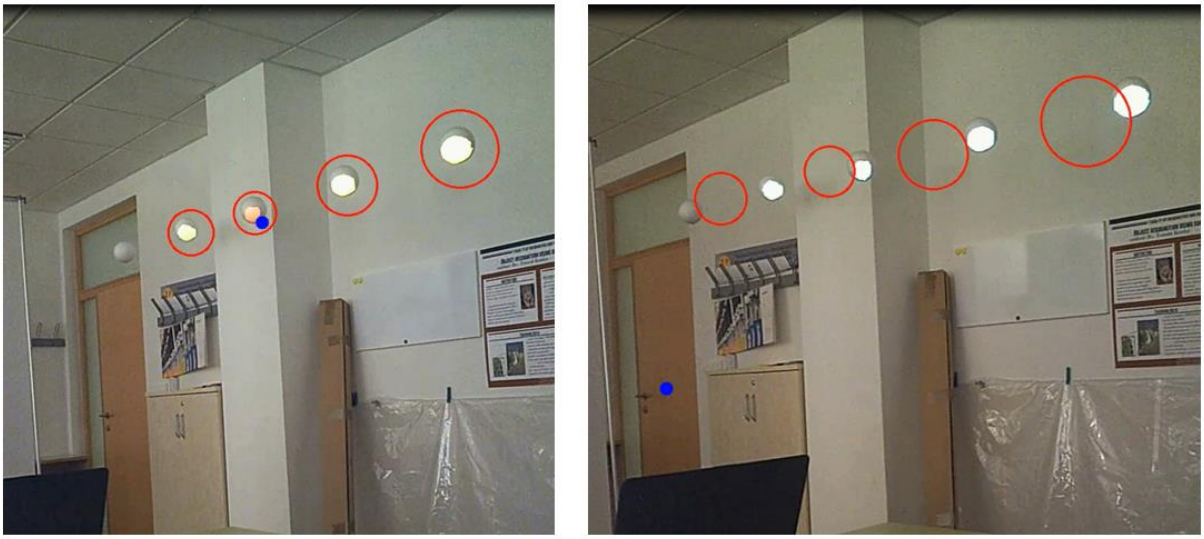


Figure 6-11: Successful tracking of previously segmented ROIs using proposed shift vector computations (left). Tracking error (right) caused by specifics of the egocentric video- constant shakes and fast tilts of the head- and by cumulation of the tracking error. The captures are from the prototyping phase of the module proposal.

6.1.3.4 Output of the automatic evaluation module

Having the information about ROI coordinates at each frame and the gaze information alongside, the detection of fixated ROI at each video frame is a trivial task. Fixated ROI at certain frame is the one intersected by the gaze coordinates. If no intersection was found, no fixation on ROI was encountered in the frame. The segmented and tracked ROIs are labeled for the evaluation output purposes from the left-hand-side (or right-hand-side if needed) starting from the zero. The fixated ROI indexes (with reference to the frame in the video) are written in the structured data file which is an output of the automatic evaluation module or are used in the graphic output of the module (as on Figure 6-11). Graphic output of the module is the egocentric video with ROI locations and gaze written in alongside with the fixated ROI index information.

We encountered significant, unpredictable, pattern-less errors in the gaze data provided by the eye-tracker (described more in detail in the Subchapter 7.1). Therefore, we take into account enlarged area around the ROIs when determining their intersection with the gaze (also on Figure 6-11). Simplified pseudocode of our proposed automatic evaluation module can be found on Figure 6-12.

Figure 6-12 Automatically evaluated egocentric video frame from the dataset by the automatic evaluation module (graphic output). Tracked ROIs are in red circles (they have enlarged areas) and gaze is a blue dot. The intersection is displayed in the top-left corner. ROIs are numbered from the right-hand-side starting from the zero.

```
video_frames <- LOADVIDEO()
roi_number <- INITROINUM()
segmentation_frequency <- INITSEGFREQ()
counter <- 0

WHILE frame <- video_frames.next()

    IF (counter % segmentation_frequency) = TRUE
        roi <- SEGMENTATION()
    END IF

    optical_flow_map <- CALCOPTICALFLOW()
    shifted_roi <- MEANSHIFT(optical_flow_map)

    roi_with_gaze <- ROIGAZEINTERSECTION(roi, gaze)

    WRITEOUTPUT(roi_with_gaze)
    counter <- counter + 1

END WHILE
```

Figure 6-13: Simplified pseudocode of the main procedure in the automatic evaluation module.

6.2 Methodology of the user study

In this chapter we will thoroughly explain the methodology of the proposed user study, making use of the user study setup described in the previous chapter, leading to achieve goals of the proposed method. As stated before, the main objective of conducting the user studies following our proposed method is creating a novel dataset for the research of visual attention. Methodology of the proposed user study is following these steps:

1. laboratory preparation
2. equipping and guidance of the participant
3. projection sequence setup
4. data storing and finalizing the dataset.

All the steps are described more in detail in the next Subchapters, so anyone can reproduce the user studies following our proposed method and user study setup.

6.2.1 Laboratory preparation

We have to ensure that the laboratory is well prepared for the experiments with human visual attention. From the theory about bottom-up and top-down factors affecting visual attention (Subchapter 2.2.1 and 2.2.2) we know that the scene observed by an experiment participant shall be as less visually distracting as possible. Therefore, saliency of the laboratory scene itself and of the objects at the scene shall be low. Hence, saliency of projection changes on the scene will be very high comparing to the neutral scene. If we ensure compliance of the laboratory setup with the proposal, we are able to study nearly isolated effect of the changes on the scene on the participant's visual attention. Low saliency of the scene can be achieved by providing e.g. white background without any distracting objects or furniture in the field of vision of the observer. This applies for the ROIs, too. Therefore, polystyrene balls are a good choice of ROIs on the scene while they are of the same colour as the background (white colour) and are not salient. Polystyrene balls should be placed at the scene following schema on Figure 5-1. It is necessary to place ROIs in sufficient distance from the background (at least 20 centimetres), so they can be segmented out from the background during the projection calibration phase. There should be low light conditions in the laboratory (not too dark, but no sunshine) and no bright objects or objects emitting light at the scene.

The second part of the laboratory preparation is projection calibration which is described in detail in the Subchapter 6.1.1 and step-by-step explained in the user guide (Appendix B of this thesis).

6.2.2 Equipping and guidance of the participant

Participant of the user study is equipped with the eye-tracker glasses, firmly tightened on the participants head to prevent shakes or moves of the eye-tracker (Figure 6-8). The eye-tracker is connected to a computer with the data acquisition module running and with enough storage space on the hard drive for the recorded egocentric video. Participants are asked to take off their dioptric glasses before participating in the experiment. The dioptric glasses would prevent the eye-tracker to work correctly because of the interferences with infrared light emitted by the eye-tracker towards the observer's eyeballs. Moreover, participants are instructed to stand still at a specified place in the laboratory and to look around the scene only with their eyes, without moving their head (see schema on Figure 5-1 and capture from the user studies on Figure 5-2). This is a preventive instruction as the head moves may cause the eye-tracker to shift and, thus, to fail in further gaze estimation.

We calibrate the eye-tracker with the participant using three-point calibration method.[7] Calibrating the eye-tracker requires three stable points on the scene far away from each other but without the need of head movement of the participant. Participant is instructed to look at them at specified order. These

[7] Explanation of three-point calibration of the eye-tracker available on 24/02/2018 at: http://tsgdoc.socsci.ru.nl/images/c/ cb/IView_X_SDK_Manual.pdf

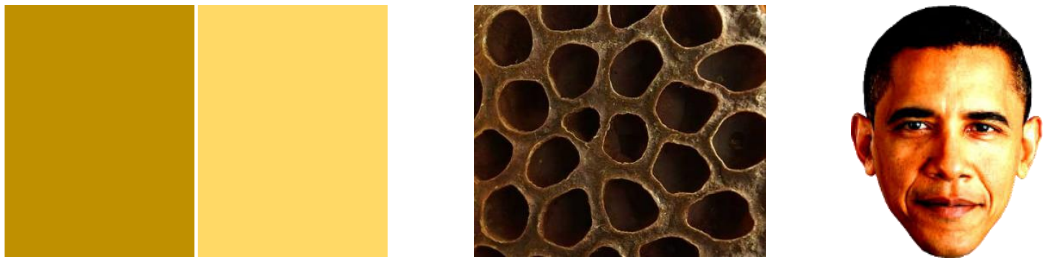
points are clicked in the stream of the egocentric video from the eye-tracking module. After the eye-tracker is calibrated with the participant, he/she is asked to look at ROIs on the scene, one-by-one, for double checking the gaze accuracy. If the gaze is not accurate enough for the further evaluation purposes, the calibration process is repeated.

After the successful calibration, participant is asked to look freely on the scene without any specific task given. By free-viewing tasks we try to suppress certain top-down factors that affect the visual attention (visual searching, previous expectations, etc.) as we want to observe mainly the bottom-up influence of the depth on visual attention.

6.2.3 Projection sequence proposal

Content projected on the scene and its changes may reflect goal of the research as the projection handling software module is open for a modification of the projection sequence. We proposed two own projection sequences for the research of depth influence on human visual attention. First proposed projection sequence was the first one, experimental, and aimed on verification of relevancy of claims about visual attention we made earlier. The first projection sequence was used to test the user study setup and software modules, and to find out if the claims about the depth influence on human visual attention are relevant to study by the proposed novel method. Moreover, we wanted to obtain the first results as an indication of possible results of depth influence on human visual attention. The first sequence consists of various changes occurring on the scene on two or more ROIs at the same to achieve concurrency aspect:

- projection of slightly different colour tint: bright yellow colour *RGB(100,100,0)* is projected on all ROIs and dark yellow colour *RGB(60,60,0)* is projected on some of the ROIs (Olešová, 2016) – Figure 6-14 a),
- projection of the same texture on two or three ROIs at the same time, as the texture projected on ROIs will be rather salient while nothing is projected on other ROIs – Figure 6-14 b),
- projection of the same face on two or three ROIs at the same time, as human face is one of the most important top-down features to attract visual attention (Xu et al., 2015) – Figure 6-14 c).



a) b) c)

Figure 6-14 Projection changes of the first proposed projection sequence. Various changes occur on the scene on two or more ROIs at the same time. Dark yellow colour is projected on certain ROIs (Figure a) left) while bright yellow colour is projected on the others (Figure a) right), texture invoking fear (trypophobia) on Figure b), or the same face is projected on certain ROIs (Figure c)).

We manually defined a few projection changes on two or three ROIs at the same time for each of the projection change types on Figure 6-14. The changes occurred on these ROIs in the specified order while other remained unaffected by the projection:

ROI depth [m]	ROI depth [m]	ROI depth [m]
2,16	6,31	-
2,68	6,00	-
3,79	5,45	-
3,23	5,72	-
4,33	4,88	-
2,16	4,88	6,31
2,16	5,45	5,72
2,16	5,45	6,00
2,68	4,88	6,31

Table 6-1 List of manually defined projection changes on two or three ROIs at the same time to test the concurrency of saliency of the changes at the scene.

Each change of the first sequence is proposed to be projected on the scene and held for 2000 milliseconds with 500 milliseconds fade-in and fade-out effect. Changes in the projection sequence defined in the Table 6-1 are not sufficient for a relevant research. However, it's implementation in the pilot user studies revealed shortfalls of the implemented software modules, laboratory setup and in the processes of the user study and resulted in fine-tuning of the proposal for more precise research. The major drawback of the sequence is the inability to mutually compare the saliency of a specific ROI in certain depth with each other. Moreover, the memory factor plays a significant role in the projection sequence parts where the same texture and face are projected because there are defined only a few changes on the scene. All the knowledge from the pilot experiments using the first proposed projection sequence was re-used to master the second proposed projection sequence which was used during the extensive major experiments with numerous participants.

The main specificity of second, enhanced projection sequence is changing of the same projected content only on two ROIs at the scene at the same time (the action is further referenced as concurrent change on ROIs). Simplified, no content is projected on the ROIs at the beginning. Subsequently, the same content is projected on two ROIs simultaneously in different depths for duration of 1000 milliseconds with 300 milliseconds lasting fade-in and fade-out effect. There are generated complete $\binom{10}{2}$ combinations of these concurrent changes on the scene as we use ten polystyrene balls and require changes on two of them at once. Therefore, during the evaluation phase, each ROI can be compared with each other, i.e. in the means of the observer's first fixation immediately after the change on concurrent ROIs occurs. Generated combination's orders are randomized to eliminate the memory factor during the user study. Three types of concurrent changes on the scene are generated in random order:

- projection of plain white colour on two ROIs on the scene (for simplicity to eliminate the influence of various aspects affecting visual attention, e.g. different colour, shape, orientation, etc.)- Figure 6-15,
- projection of slightly different colour tint: bright yellow colour *RGB(100,100,0)* is projected on all ROIs and dark yellow colour *RGB(60,60,0)* is projected on two changed ROIs (Olešová, 2016) – Figure 6-16,
- projection of the same face on two ROIs, as human face is one of the most visually attractive objects known these days (Xu, 2015)- Figure 6-17.

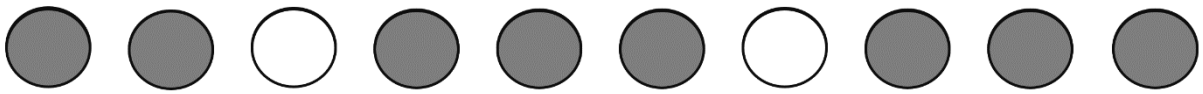


Figure 6-15 Visualization of projection of plain white colour on two random ROIs.

Figure 6-16 Visualization of projection of slightly different colour tint on two random ROIs.

Figure 6-17 Visualization of projection of the same faces on two random ROIs.

Projection sequence rendered by the projection handling module is ready to be played on one click right after the participant undergoes the successful calibration process. The sequence has to be played as a whole, without any interruption or interaction with the projection handling module.

6.2.4 Data storing

The egocentric video file and corresponding data file with gaze information from the user study are automatically stored in the unique folder created by the data acquisition module for each participant. The file formats and principles of data storing is described in the Subchapter 6.1.2.

The data acquisition procedure has to be finished properly as described in the instructions printed by the eye-tracking module (for more details see the User guide in the Appendix B of this thesis). This is due to necessity of proper finalization of the created video file. Otherwise, the video file may be corrupted and not usable for evaluation. Captured data are ready to be processed and evaluated either by the researcher, or automatically using our proposed data evaluation module.

6.3 Technical and implementation details

Windows 10 with the latest updates was the platform we used for implementation and use of the proposed modules. Selection of platform was dependent on the requirements of Kinect 2.0 device which operates well on Windows platforms. Microsoft Visual Studio 2015 was used as our development environment.

We implemented proposed software modules in C++ language. We used several library dependencies to fulfil implementation needs of the modules. For image processing, computer vision and video manipulation purposes we used the open-source self-compiled OpenCV library in version 3.4[8] together

[8] Available on 01/05/2018 at: https://opencv.org/opencv-3-4.html

with the optional contributions package[9]. We used the open-source GLFW 3 library[10] for operations on graphic chip of the computer required by the projection handling module.

Setting up the acquisition of the data from Kinect 2.0 device and its integration to a custom C++ project was more complex, as the data from the Kinect should be in format compatible with the OpenCV library for further processing. We installed Kinect for Windows Runtime v2.0 driver and Kinect for Windows SDK v2.0[11]. For testing purposes of the Kinect device, we installed the Kinect Studio v2.0, too. The platform for interaction with the Kinect 2.0 SDK and transformation of the outputs to the OpenCV format is a piece of work of Yoshihisa Nitta released under MIT license for educational purposes[12].

The SMI SDK is not currently publicly available as the SMI company exists no more since 2017. The former SMI webpage provides no longer any data or support. We obtained the SDK without any documentation from the master thesis of Olešová (Olešová, 2016).

We implemented each software module as a separate C++ project with possible release of independent binaries on each other. Main parts of the modules from the technical and implementational point of view are described in the Appendix A of this thesis.

[9] Available on 01/05/2018 at: https://github.com/opencv/opencv_contrib
[10] Available on 01/05/2018 at: http://www.glfw.org/
[11] Available on 01/05/2018 at: https://developer.microsoft.com/en-us/windows/kinect/develop
[12] Available on 01/05/2018 at: https://github.com/YoshihisaNitta/NtKinect

7 Evaluation and results

We describe conducted user studies, results from the created datasets and their evaluation in this Chapter. We held two experiments testing and verifying our software modules and user study setup and two experiments (pilot and major user studies on depth impact on the human visual attention) following proposed methods and methodology in this thesis (Chapters 5 and 6). The last two mentioned experiments (pilot and major one) were involving complete user study setup and the projection sequences defined in the Subchapter 6.2.3.

7.1 Testing the setup during first experiments

Before our own user studies of the depth impact on visual attention took place, we evaluated proposed software modules implementations by numerous tests and two small experiments that took place at our faculty. These experiments served us as a proof of concept of our method proposal and as a feedback on the software modules, so these could have been refined for our major user studies. The two test experiments were:

- participation on the user studies of emotion impact on the human visual attention (as a result of collaboration between the Faculty of Psychology, Comenius University in Bratislava and colleagues from Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava),
- our own, simplified experiment testing the modules integration.

7.1.1 User studies of emotion impact on the visual attention

We participated on the user studies of emotion impact on the human visual attention. The user studies were result of collaboration between the Faculty of Psychology of Comenius University in Bratislava and the Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava. The main objectives of our participation were to beta-test the crucial part of our proposed eye-tracking module and to provide support for the eye-tracking phase of the experiments proposed by colleagues from the faculties. One half of the eye-tracking part was held with the SMI eye-tracking glasses using SMI SDK and our proposed eye-tracking module and the other half was held in a laboratory equipped with displays with attached, static eye-trackers. We recorded egocentric videos along with gaze data from seven participants during the experiments. Example of the egocentric frame from the eye-tracker is on Figure 7-1.

During the evaluation, we encountered significant accuracy error of the gaze data from the eye-tracker. Even after precise, repeated calibration, the device was not able to measure the gaze accurately enough to distinguish certain point of fixation of visual attention. The gaze was skewed significantly during whole experiments with the gaze error reaching more than 100 pixels in the egocentric video. The gaze error was unpredictable, without specific patterns (e.g. shift of the gaze 50 pixels to the left), making it impossible to deal with the error in the post-processing phase. Moreover, the distance of the observer from the projection plane and size of the objects made the error of the gaze even more severe and in

some cases crucial for the evaluation (as on Figure 7-1). Hence, only limited manual evaluation of the obtained dataset was possible.

We have to emphasize, that the gaze error was not caused by our proposed eye-tracking module as the skewed data were the output of the SMI SDK's server-side which is closed for modifications or extensions. As we had to work with the eye-tracker during our further research, we tried to eliminate the gaze error by proper eye-tracking conditions and some client-side software modifications.

Figure 7-1: Egocentric video frame with gaze written in (red dot). The egocentric video was captured throughout the user studies of emotion impact on the visual attention held at the faculty. The participant just found searched object, located approximately in the middle of the displayed image. There is a significant gaze error encountered, here, and the evaluation of the frame is not possible.

7.1.2 Module testing during the simplified experiments

We decided to conduct small test experiment during the implementation phase of our work. Our decision was motivated by high complexity of the proposed modules and the user study setup. Therefore, we wanted to beta-test them to verify our approach, test implementations of the modules and to collect information necessary for improvement of some crucial parts of the modules, as we wanted to obtain the best results possible during the major experiments.

The experiment was held in the same laboratory where our future major experiment took place. We involved 5 participants in the experiments. Setup of the experiments had these constraints comparing to the complete proposal in this thesis:

- there was one projector used,
- projection was limited to four ROIs on the scene,
- salient objects were not removed from the background,
- simplified projection sequence was used (a few changes on the single ROIs without the concurrency aspect).

The evaluation of the module implementations is an intermediary step towards conducting our major experiments. Therefore, we took testing notes and improvement proposals which are already implemented in the setup proposal in this thesis, rather than making statistical evaluation of performance of the involved modules.

7.1.2.1 Projection handling module

Projection mask applied on the calibrated LCD projector correlated with the ROI positions at the scene. The projection itself was of satisfying quality to simulate changes at the scene and project textures (e.g. human face) without affecting the background (capture on Figure 7-2). The improvements to the calibration module were made, namely the re-calibration step proposed in the Subchapter 6.1.1, to maximize the surface of the ROIs covered by projection. There was a significant limitation of the calibration precision by the Kinect 2.0 depth map accuracy. We were also limited by avoiding more complex geometric transformations during the calibration phase, so we could stay focused on the primary goals of this thesis. We decided to implement the user interaction possibility in the module. Our aim is to provide the user a possibility to fine-tune the projection manually after the calibration phase, instead of introducing novel methods for projection calibration with higher precision. The user interaction allows to match the projection with the ROI surface perfectly- for more details see the Subchapter 6.1.1.

Figure 7-2: Successful projection calibration during tests of projection calibration module (simplified experiments).

7.1.2.2 Eye-tracking module

The SMI SDK extension proposed in this thesis was reliable with no encountered loss of the data from the eye-tracker or frame-rate drop in the egocentric video. However, as stated in the previous Subchapter, problematic part of the module was the eye-tracking hardware and server-side of the SDK. We had several calibration issues related to the SDK with the eye-tracker. The client side of the eye-tracking module itself does not have any control over calibration procedure after sending calibration data to SDK server. The calibration procedure on the server-side was, however, not reliable. It was not possible to calibrate the eye-tracker with significant number of participants: 2 out of 5 in our test experiment. The problem persisted even after several calibration attempts, making the calibration successful only in approximately 60%. The inability to calibrate the eye-tracker with a participant is making it impossible for him to participate in the experiment. Moreover, even after successful

calibration, the data were frequently skewed and not accurate enough. In some cases, it was not possible to determine on which object participant fixed his attention. This made significant problems not only for the automatic evaluation module expecting the accurate gaze data from the eye-tracker (with small tolerance) but also for manual evaluation of the data. The gaze error of the eye-tracker can be clearly seen on Figure 7-3.

Figure 7-3: Gaze error from the eye-tracker after its proper calibration. The error is unpredictable (as stated on Figure 7-1), and its causes are not clear, yet. However, the eye-tracking hardware is significantly out-dated, nowadays.

The error of the eye-tracker is more significant when working with various depths of the scene. While the gaze error in a video frame can be only few pixels, the same error represents a few centimetres when looking at distant objects (Figure 7-4).

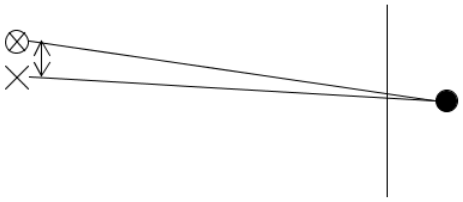


Figure 7-4: Gaze error during eye-tracking of fixations at objects in certain distance from the observer. The more far the object is, the bigger the negative impact that the error has on the eye-tracking precision is. Black dot represents the observer from the top-down view, vertical line represents the egocentric video plane and two crosses represent real objects captured in the video. One of the object (cross in the circle) represents the one where the observer really looked at and the other is the skewed gaze captured by the eye-tracker.

The gaze error is known issue of many mobile eye-tracking devices and manufacturing companies are dealing with this issue for a long time. However, the gaze error of the SMI eye-tracker is outstanding as the hardware is significantly out-dated, nowadays. Gaze data are determined from the internal eye modelling based on the shape of the eye, cornea, location of the fovea and reflectiveness of infrared light emitted from the eye-tracker. When eyes of a participant does not meet the requirements of the eye-tracker on anatomical shapes or reflectiveness, then the gaze of participant cannot be determined. This

can be avoided by modern technologies and driver improvements which represent in most cases a secret of the manufacturer. More about eye-tracker calibration can be found in an article issued by Tobii laboratories[13].

There are two partial gaze error solutions that we propose. The first one was already implemented in the automatic evaluation module. It is based on enlarging ROI detected in egocentric video by a constant error radius. The error radius helps the evaluation module to deal with fixations that seems to be out of the ROI, but in fact they are (Figure 7-5). The enlargement of the ROI locations is limited to the size that ensures no overlap of the ROIs.

The second partial solution we propose is to maintain prefect conditions during the calibration eye-tracking phase of the experiments. We were able to increase success rate of the calibration of the eye-tracker with the participants to nearly 100% by certain hardware adjustments. However, it is hard to solve the mentioned gaze error programmatically. Therefore, we suggest using mobile eye-tracker of a higher quality during the future research.

Figure 7-5: Partial solution for the eye-tracker's gaze error which was additionally proposed and implemented in the automatic evaluation module. The ROI areas (red circles) are enlarged when searching for the gaze-ROI intersection by the evaluation algorithm.

7.1.2.3 Data evaluation module

We tested the implementation of the data evaluation module as proposed in this thesis. As we mentioned in the Subchapter 6.1.3, using dense optical flow without any modification is not enough for object tracking in the egocentric video, as this problematic is a research topic itself. Anyway, using dense optical flow ends up by error cumulation from the tracking phase and, thus, maintaining inaccurate ROI locations in the egocentric video. Therefore, we proposed to repeat the segmentation phase every *n*-th video frame to adjust the ROI locations in the egocentric video and eliminate the error.

Moreover, we are often unable to run our proposed segmentation algorithm to detect ROI locations in a video frame due to projection of various content (e.g. dark colours). Therefore, bright colour has to appear on each ROI at the same time every *n* seconds, making it possible to run a segmentation step in video frames. We propose to project white colour on each ROI after each projection change at the scene,

[13] Available on 01/05/2018 at: https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/what-happens-during-the-eye-tracker-calibration/

or at the frames marked as segmentation ones in the data file with gaze information. This can serve as a resetting factor for observer's visual attention during the experiments and as a synchronization for the data evaluation module with projection sequence, too. The proposed method was implemented and was proved as working in the pilot experiments described in the next Subchapter. However, due to the accuracy error of the eye-tracker, making automatic data evaluation possible only in rare cases, it was not used during our major experiments.

7.2 Pilot experiments

The first experiments with the depth influence on human visual attention were pilot experiments with 7 participants using the first proposed projection sequence. These were held to test all the equipment and implemented software modules and to find out if the claims about the depth influence on human visual attention can be studied using the dataset created by the proposed novel method. Order of fixations of participants after each projection change at the scene was extracted from the data collected throughout the experiments. We encountered again a significant error in the gaze data obtained using the SMI eye-tracker (as stated in previous Subchapters). Therefore, automatically evaluated data were in most cases corrected by human interaction and some of them were evaluated manually. The reason is visualized on Figure 7-6.

Figure 7-6 Captures of the egocentric video frames (with gaze written in as a red dot) at the same moment during the experiments with two different participants. The capture on the left side visualizes decent quality of the obtained gaze data while the capture on the right side visualizes bad quality of the gaze data. Both participants are looking at the closer ROI with projected face.

We are providing visualizations of participants first fixation ratio after projection change at the scene occurred (Figures 7-7 – 7-10). First fixations ratio is expressed in percentage and grouped by three types of the projection sequence changes (see the first projection sequence proposal in the Subchapter 6.2.3). We can discuss, regarding the obtained data, that distribution of ROI's saliency in similar depths is approximately evenly distributed (Figure 7-8). On the other hand, distribution of saliency of ROIs which depth is significantly different is not balanced (Figure 7-7) and invokes an assumption that objects closer to the observer are more salient. The findings and first results prove our first claim that depth plays significant role as an aspect of human visual attention and further research of depth influence on the visual attention is relevant.

We can observe from the first results that first fixations of participants may be in relation with the type of change at the scene (either change in colour, texture presence or face presence). Slight changes in colour were surprisingly more salient in greater distance from the observer in comparison to texture or

face projection changes in same distances. This fact, however, may be only a coincidence caused by small dataset. Face and texture changes on ROIs were significantly more salient in distances closer to the participant when changes on two ROIs occurred. This can be clearly seen on Figures 7-9 and 7-10. Concurrency of changes on three ROIs proved the assumption that any change closer to participant was the most salient one and the saliency of changes on the scene were decreasing with the distance from the observer. This can prove our second claim that depth influence on the human visual attention can be approximated as a continuous function (Olešová, 2016) and the exact relation of the depth influence and the saliency is relevant for further research. Moreover, as stated before, saliency of the same objects in similar distances from the observer are approximately evenly distributed and balanced (Figure 7-9).

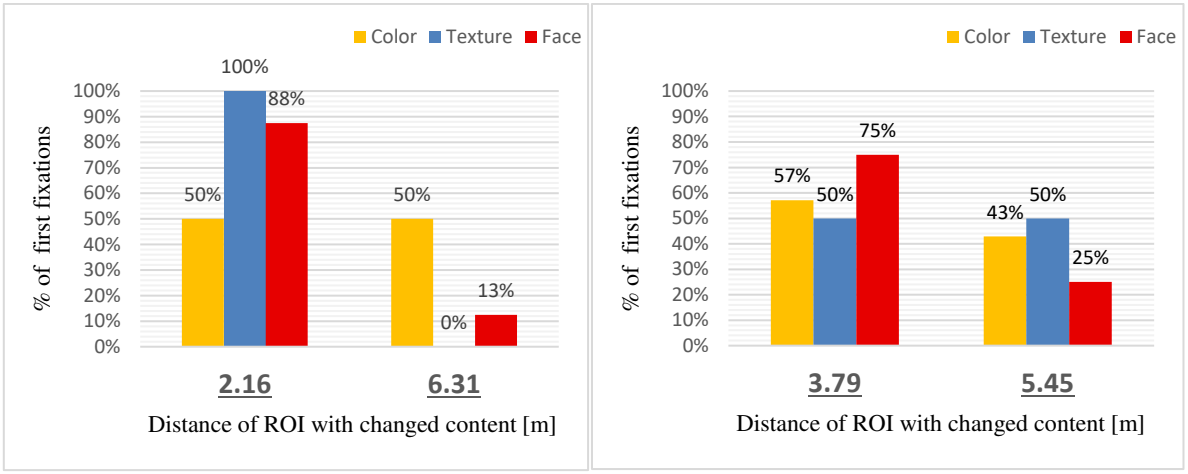


Figure 7-7 First fixations ratio after concurrent changes on two ROIs at significantly different depths occurred. Type of change is expressed as colour of the column in the bar chart. Generally, we can observe that ROI closer to the observer was more salient. Moreover, type of the change on the scene may correlate with its objective saliency which is indifferent on the depth of the scene.

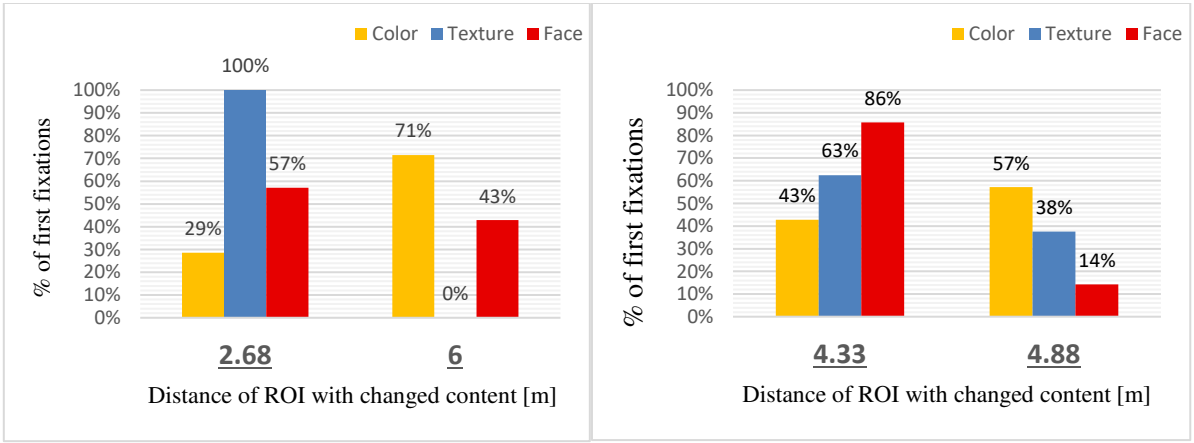


Figure 7-8 First fixations ratio after concurrent changes on two ROIs at similar depths occurred. Generally, the distribution of first fixations is more balanced. However, the aspect of the type of change at the scene is very strong.

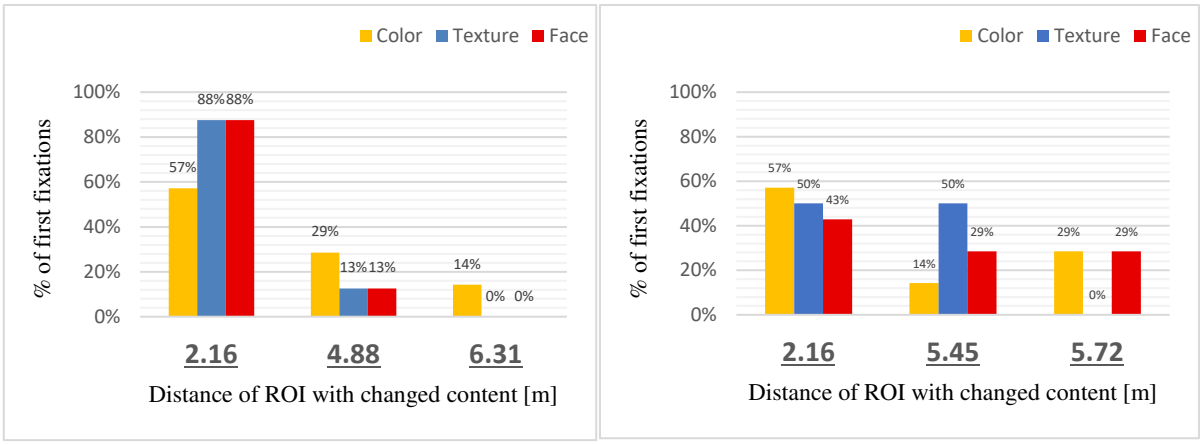


Figure 7-9 First fixations ratio after concurrent changes on three ROIs occurred. One ROI was in the depth close to the observer and the other ones were in different, further depths. Saliency in the diagram on the right is more evenly distributed than in the right one. This may be the consequence of similar depths of two of the ROIs affected by the projection in the results provided by the right diagram.

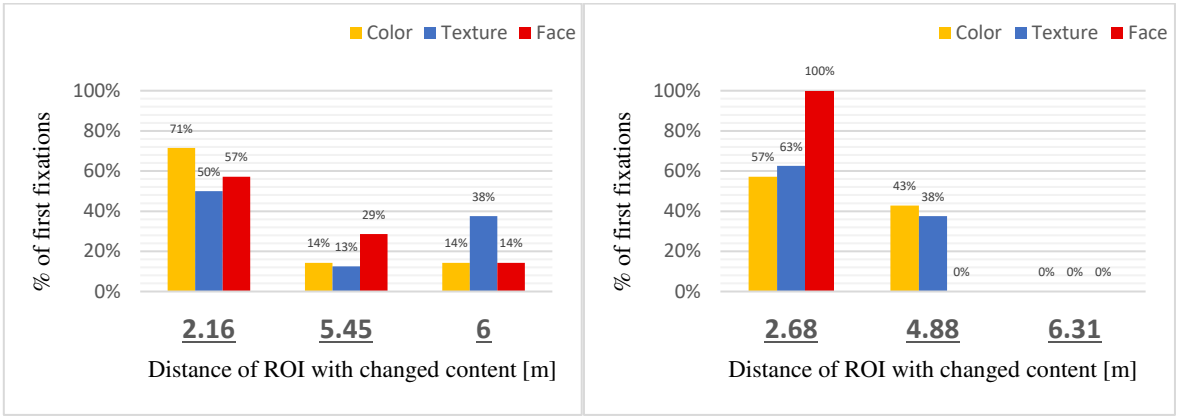


Figure 7-10 First fixations ratio after concurrent changes on three ROIs occurred. One ROI was in the depth close to the observer and the other ones were in different further depths. It can be clearly noticed, that the ROI closest to the observer attracted the observer's attention in most cases.

In addition, we have to mention that results of first fixations ratio on Figures 7-7 – 7-10 may be influenced by memory factors. The changing colour tone, texture and projected face could have been memorized from occurrences at the scene before, as the order of projection sequence was corresponding with order in the Table 6-1. This means that top-level factors might have influenced participant's visual attention, and these were dominating instead of bottom-up factors in our results. Moreover, the dataset is too small to conclude any further claims from the findings summed up in the previous paragraphs. We tested our software modules which were reliable and identified shortfalls of the hardware we are using (mainly the accuracy and calibration performance of the eye-tracker). The precision of the eye-tracking hardware is a true limitation for the user studies as the hardware is very specific, expensive, and cannot be easily replaced.

We revised the module proposals to even better match the requirements and mastered the projection sequence to better match our research goals in research of depth as the aspect of human visual attention. Summing the results up, we obtained the indicators saying that our research and our claims are relevant, and we obtained first results of depth influence on the visual attention which we expect to be proved in the extensive major experiments.

7.3 Major experiments

We held the major experiments with the same laboratory setup as proposed in the thesis and verified in the pilot experiments. The only difference was the refined projection sequence which was used (the second one mentioned in the Subchapter 6.2.3). There was big amount of the data obtained throughout the major experiments that took place during three full days in the laboratory. The experiments were more exhausting and time-consuming comparing to the pilot ones while the projection sequence consisted of three parts with complete combinations of concurrent changes at the scene. We also did our best to calibrate the eye-tracker with each of the participants as thoroughly as possible to produce a novel dataset of a high quality. Time spent with one participant was, therefore, approximately 15 minutes of which approximately 8 minutes were related to monitoring their visual attention while the projection sequence took place.

We created the novel dataset (contents described in the Section 5.1.4) from the data obtained by conducting the user studies with 37 participants. Part of the information relevant for our thesis and carried by the dataset was evaluated during studying the depth influence on human visual attention. We evaluated observer's first fixation on certain ROI immediately after projection change at the scene occurred. Other information about visual attention of the observer (i.e. duration of the first fixation, its delay after the change on the scene) are not considered, yet. They are too complex for evaluation and can be included in the future research. Moreover, we take into account only the first part of the projection sequence involved in the experiments (projection of a white colour on two ROIs at the same time). This evaluation approach is based on findings from the pilot experiments where more complex changes on the scene like colour tint change or face presence means significant influence of other aspects on the observer's visual attention. These aspects were interfering with the aspect which is subject of our research- depth of the objects at the scene. Further parts of the projection sequence were included in the experiments to provide more extensive dataset for possible future research of the human visual attention from the egocentric perspective of view.

During the evaluation process, each ROI was compared with each other in the means of the first fixation ratio after the concurrent change on the scene occurred (as proposed in the Subchapter 6.2.3). The ratio is expressed as a percentage value. Visualizations of the comparison of the all the ROIs in certain depth with every other ROI are on Figures 7-11 – 7-15. There can be clearly distinguished more salient ROIs with comparison to the other ones (e.g. ROI in the depth of 4.24m with very high fixation ratio comparing to each other ROI).

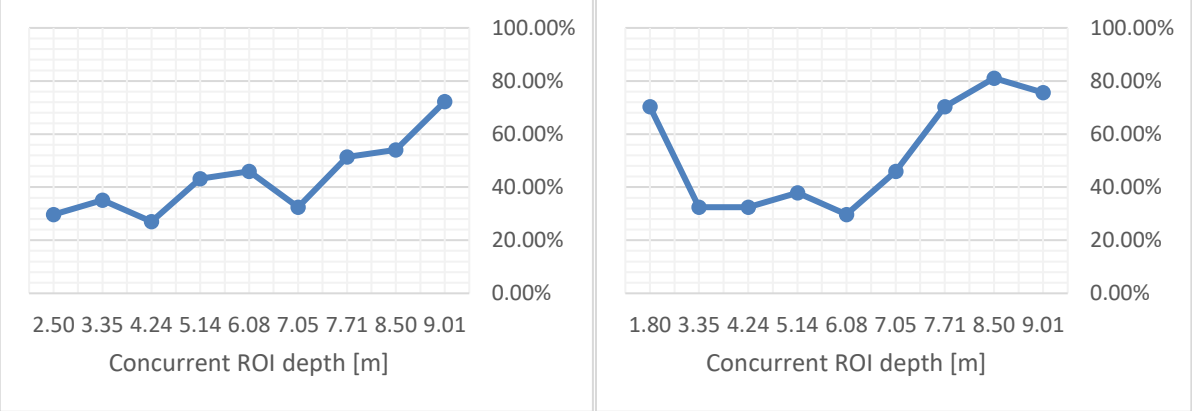


Figure 7-11 First fixations ratio (in percentage) on the ROI in the depth of 1.80 metres (left) and 2.50 metres (right) in comparison with every other ROI after concurrent change on the scene occurred.

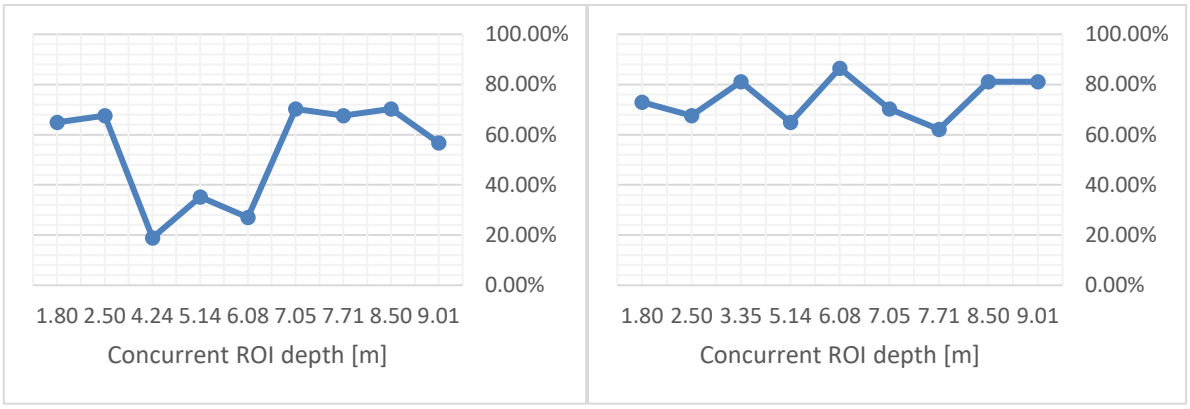


Figure 7-12 First fixations ratio (in percentage) on the ROI in the depth of 3.35 metres (left) and 4.24 metres (right) in comparison with every other ROI after concurrent change on the scene occurred.

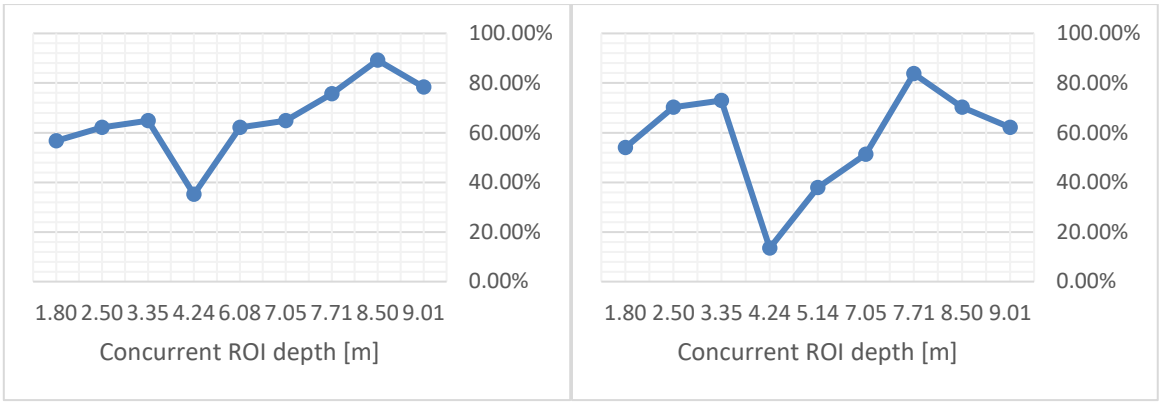


Figure 7-13 First fixations ratio (in percentage) on the ROI in the depth of 5.14 metres (left) and 6.08 metres (right) in comparison with every other ROI after concurrent change on the scene occurred.

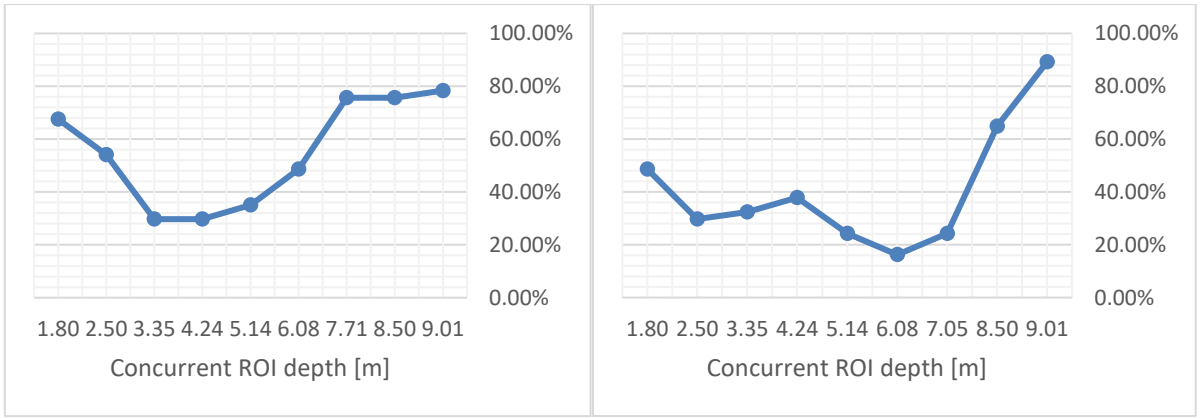


Figure 7-14 First fixations ratio (in percentage) on the ROI in the depth of 7.05 metres (left) and 7.71 metres (right) in comparison with every other ROI after concurrent change on the scene occurred.

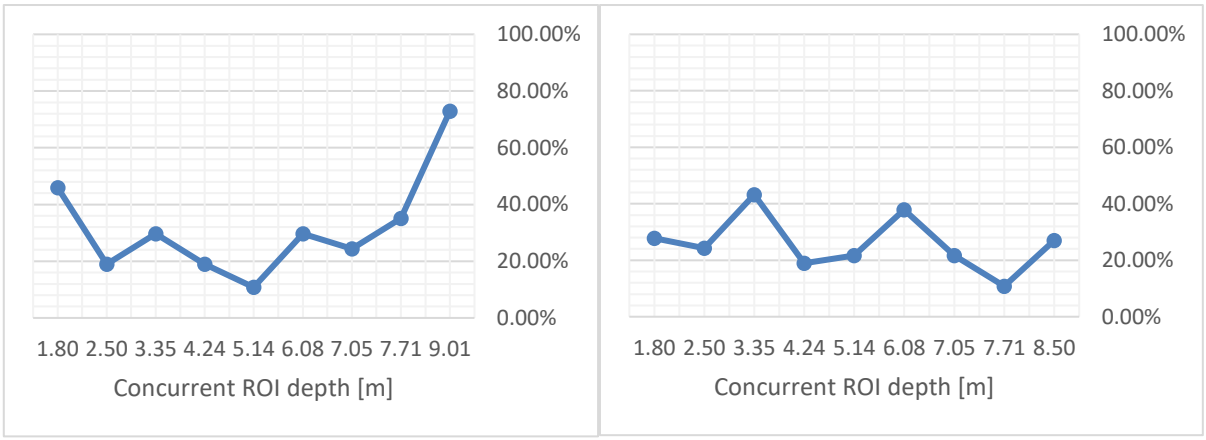


Figure 7-15 First fixations ratio (in percentage) on the ROI in the depth of 8.50 metres (left) and 9.01 metres (right) in comparison with every other ROI after concurrent change on the scene occurred.

Considering that concurrent changes on two ROIs are evaluated in the statistics we claim that the first fixations ratio of 50% refers to a pure chance (each of the two ROIs have attracted the same amount of first fixations of the observers). Thus, we cannot conclude anything about the relation of the ROI's saliency and the distance of the ROI from the observer. Subsequently, we claim that any percentage of first fixations over (or below) 50% means that there may be a relation of the ROI's saliency and the distance of the ROI from the observer. In the case of percentage value of the first fixations over 50% we speak about a positive relation: saliency of the ROI in certain depth is higher in comparison with the other one in different depth. On the other hand, in the case of percentage value of first fixations below 50% we speak about a negative relation- saliency of the ROI in certain depth is lower in comparison with the other one in different depth. Therefore, we introduce our own metrics- depth score- for determining impact of the ROI depth on its saliency. The metric is used for our own statistical and evaluation purposes and to support some of our claims. Depth score is simply defined as a fixation ratio's percentage value normalized to the interval *<0;1>* from which the normalized value of a pure chance is subtracted or added:

$$score' = norm(percentage);\ score' \in\ < 0; 1 >$$

$$score = score' \pm 0{,}5\,.$$

The operation of subtraction or addition depends on the desired range of the depth score and on the visualization or application the score is used for. The subtraction shuffles range of the depth scores to the interval *<-0,5;0,5>*. This range is good for visualization of collision table, comparing depth score of every ROI with each other (Table 7-1).

meters	1,80	2,50	3,35	4,24	5,14	6,08	7,05	7,71	8,50	9,01
1,80		-0,20	-0,15	-0,23	-0,07	-0,04	-0,18	0,01	0,04	0,22
2,50	0,20		-0,18	-0,18	-0,12	-0,20	-0,04	0,20	0,31	0,26
3,35	0,15	0,18		-0,31	-0,15	-0,23	0,20	0,18	0,20	0,07
4,24	0,23	0,18	0,31		0,15	0,36	0,20	0,12	0,31	0,31
5,14	0,07	0,12	0,15	-0,15		0,12	0,15	0,26	0,39	0,28
6,08	0,04	0,20	0,23	-0,36	-0,12		0,01	0,34	0,20	0,12
7,05	0,18	0,04	-0,20	-0,20	-0,15	-0,01		0,26	0,26	0,28
7,71	-0,01	-0,20	-0,18	-0,12	-0,26	-0,34	-0,26		0,15	0,39
8,50	-0,04	-0,31	-0,20	-0,31	-0,39	-0,20	-0,26	-0,15		0,23
9,01	-0,22	-0,26	-0,07	-0,31	-0,28	-0,12	-0,28	-0,39	-0,23	

Table 7-1 Collision table of the depth scores based on the visual attention data collected during the extensive user studies with 37 participants. The table should be read from the left to the right. There is a strong trend in the depth score data forming the illusionary diagonals. Thus, depth perception patterns are proved to exist and they form strong aspects influencing the visual attention.

We can make some conclusions based on the analysis of the provided visualization. The most important conclusion is that depth plays definitely a significant role as an aspect of the human visual attention as there are strong patterns of depth visual perception based on visual analysis of the collision table. We can prove this claim by the fact that the values of the depth scores do not converge to zero and vary in the whole interval $\langle -0,5;0,5 \rangle$ and there is a strong trend in the depth scores, forming the illusionary diagonals.

We can observe the trends in the score values from the visualization. Depth scores under the diagonal from the top-left corner are positive, looking at them from the left-hand-side to the right-hand-side. This means that the objects closer to the observer are more salient in relation to their depth than the distant ones. In other words, the saliency of the objects at the scene was decreasing with their distance from the observer. However, this is not true about the objects too close to the observer. This can be clearly seen by scores of the ROIs too close to the observer, namely the first three one at depths of 1,80 m, 2,30 m and 3,35 m. They were more salient than the farthest ROIs, but less salient as less distant ROIs farther from them.

This leads us to an assumption, that human visual attention assigns the highest saliency to objects in certain depth at the scene. We introduce this phenomenon as the most salient depth. The saliency of the objects is then decreasing with the distance of the objects from the most salient depth. This assumption is supported by Figure 7-16 displaying the average scores of all the ROIs in different depths computed from the collision table (the average values are normalized to the interval $\langle 0,5;1,5 \rangle$).

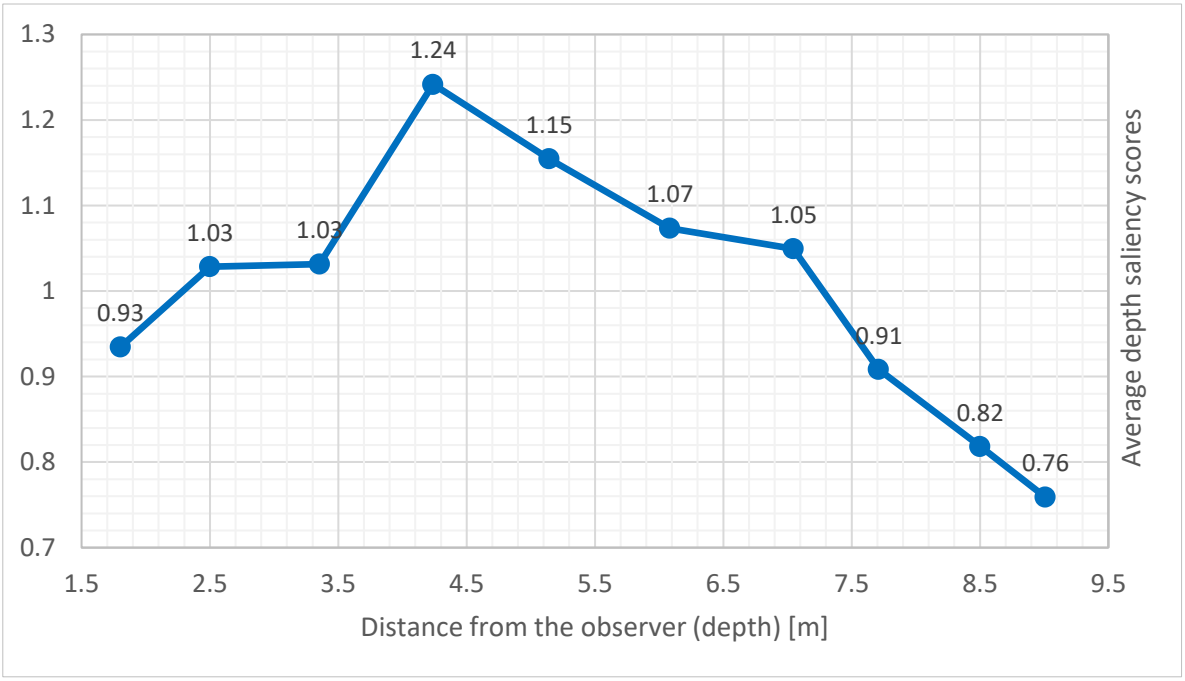


Figure 7-16 Average depth saliency scores computed as the statistical mean of ROI scores against each other ROI at various depths (visualized in Table 7-1). The most salient ROI and, thus, the most salient depth was the one in approximately 4.24 metres from the observer. Saliency of the objects in other depths were significantly decreasing with the distance from the most salient depth. The scores are normalized to the interval $\langle 0.5;1.5 \rangle$.

7.4 Depth influence on human visual attention

The average depth saliency scores on Figure 7-16 carry information about the statistically average saliency of the object in certain depth based on the ground-truth data. The average depth difference between ROIs was 72.1 centimetres which refers to sampling frequency of the average depth score curve on Figure 7-16. The connected curve forms the basis of our novel depth saliency coefficient approximation. The approximation can be derived by defining a continuous, connected function converging to the ground-truth of the observed data. Depth saliency coefficient expressed by a continuous function is then applicable as a weighting factor on the existing saliency models when the information about depth of the scene is available. The conversion of our findings about the depth influence on human visual attention from the egocentric perspective to the camera perspective of the existing saliency models is the ultimate goal of our research. It connects our novel proposed approach to visual attention research with the common, traditional approaches meaning more significant contribution to the area of visual attention modelling.

The approximation of general depth saliency coefficient based on the ground truth of obtained data should be a connected function defined in the range of $(0; \infty)$ meters where the infinity is just a mathematical generalization. In praxis, the range is upper-bounded by human sight range in meters which varies from person to person. Research work of Olešová (Olešová, 2016) concludes approximation of depth influence on the human visual attention as a polynomial function of the third order (see Subchapter 4.1.2). This representation has many limitations among which the biggest are: small range of depths on which the function can be applied (the range is approximately $\langle 0.5; 3.5 \rangle$ meters) and low accuracy of the depth influence as the polynomic curve ignores the peaks in the visual attention ground truth.

We decided to approximate introduced depth saliency coefficient as a conditional continuous function. This representation was chosen to preserve significant peek in the range of the most salient depth in the ground truth data and to better reflect gradual descend of the coefficient in relation with the distance from the most salient depth. We choose two exponential functions with Euler number as a basis with intersection in the most salient depth of 4.24 metres to model the slopes on both sides of the most salient depth. The exponential approximation of the ground truth data is visualized on Figure 7-17 as the data trendlines and the whole function is visualized in wider range on Figure 7-18.

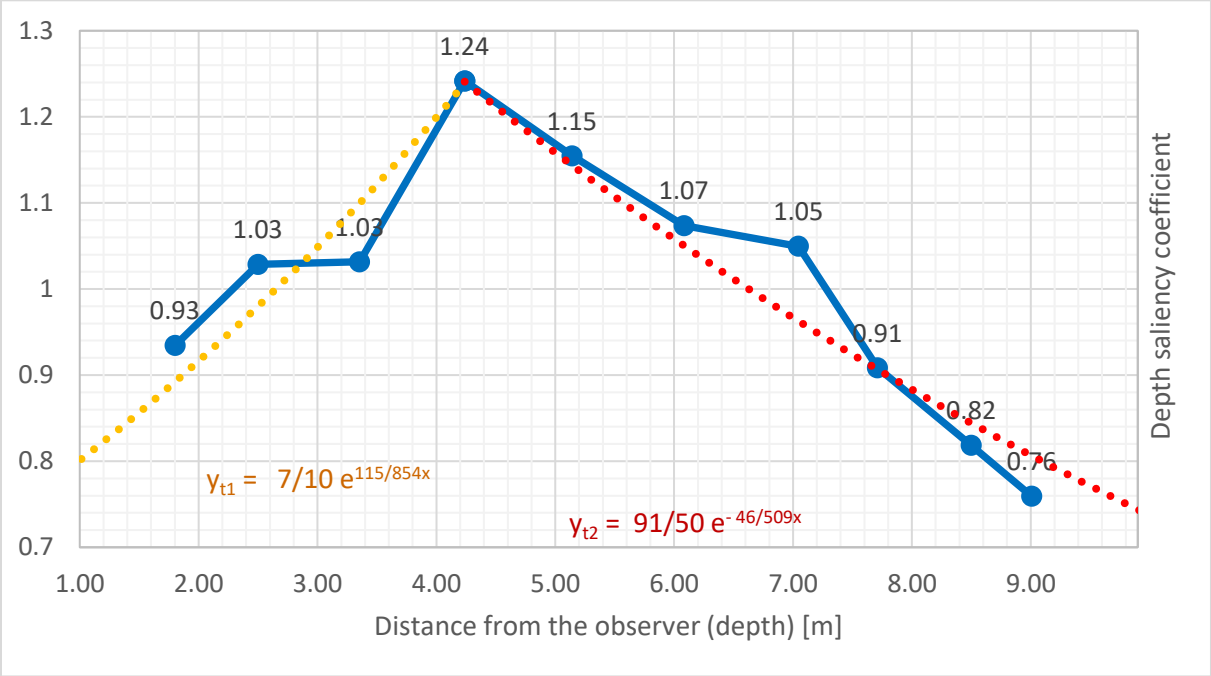


Figure 7-17 Depth saliency coefficient normalized to the range *<0.5; 1.5>* with the trendlines of continuous conditional function approximating the ground truth of visual attention data obtained during the user studies.

We can define a connected continuous conditional function of the depth saliency coefficient using the equations of the exponential functions approximated on Figure 7-17 above. We introduce the depth saliency coefficient function as:

$$y = \begin{cases} \frac{7}{10} e^{\frac{115}{854}x}, & if\ x \in (0; 4.24) \\ \frac{91}{50} e^{\frac{-46}{509}x}, & otherwise \end{cases}$$

where x is the depth in meters and y is the depth saliency coefficient relative to the depth. The depth saliency coefficient function is visualized on Figure 7-18 in the range *(0; 60>*. It is obvious that we made a step forward in the depth-saliency function definition and approximated the depth-saliency coefficient for further depths than the observed ones during the major experiments. We have to emphasize that the function trend in the further depths is only an assumption and a forecast. Anyway, we overcame the shortfall of the depth range limitation of depth influence on visual attention as we had enough data to make the forecasting possible. Moreover, we assume that the trendline should possibly match the influence of depth on visual saliency even in more further depths while the trend of the function seems very promising. The function converges to zero from the depth of approximately 60 metres which is a veracious assumption that can be proved in the real scenes with great variety of depths in the future.

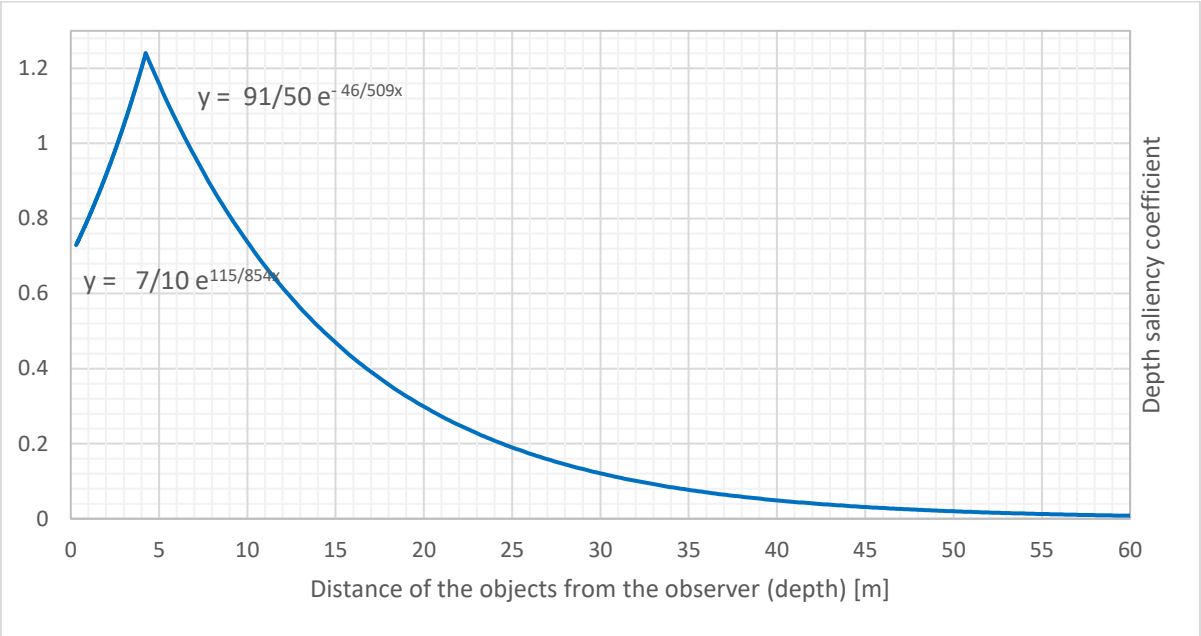


Figure 7-18 Visualization of the most important portion of the proposed depth saliency coefficient function. The function definitions are listed below the trendline (for the interval *(0;4.24>*) and above the trendline (for the interval *(4.24; ∞)*). Notice, that the function converges to zero from the depth of approximately 60 metres, reflecting the expected coefficient function behaviour.

7.5 Depth saliency coefficient evaluation

We evaluated proposed depth saliency coefficient applied on the existing saliency model proposed by Olešová (Olešová, 2016). We chose this model because of similar approach of depth incorporation into the saliency modelling as we propose in this thesis and because of comparison possibilities with previous related work. We incorporated our depth saliency coefficient to the existing model with depth contrast by:

- weighting the depth contrast coefficient of the superpixel by depth saliency contrast using mean superpixel's depth (model further referenced as DC+SCC)
- weighting each saliency map value by depth saliency coefficient using depth information at each point of the image (model further referenced as DC+SC).

For each of the depth saliency coefficient incorporation methods we created a novel saliency model. There were three other saliency models proposed by Olešová that were used for results comparison (all described in Olešová, 2016):

- model with no depth information incorporation (model further referenced as "no depth"),
- model with depth contrast only (model further referenced as DC),
- model with depth contrast and weighting of each saliency map value by the coefficient of depth influence on the visual attention, proposed by Olešová, using depth information at each point of the image (model further referenced as DC+OC).

We evaluated these models on the complete NUS-3D saliency dataset (Lang et al., 2012). More details about the dataset are provided in the Subchapter 4.1.3. The software module implemented for the evaluation purposes by Olešová (Olešová, 2016) was reused and modified for our model evaluation needs. Modification included support for whole dataset evaluation at once, computation of evaluation statistics over the complete dataset, integration with software module gSLICr for SLIC superpixel computation on graphic chip proposed by Birkus (Birkus, 2015) and corrections of evaluation formulas (general ROC/AUC formula replaced by the more widely used AUC-Judd (Judd et al., 2012)). Saliency map was generated by each of the mentioned models for each of the 600 images with real scenes in the dataset accompanied with the depth maps captured by Kinect device. The ground-truth images with fixation information for each 3-D image in the dataset were used to measure the evaluation metrics described in the following subchapter.

7.5.1 Evaluation metrics

We evaluated the visual attention models by a few common metrics for saliency model evaluation. We chose our evaluation metrics according to the paper by Bylinskii (Bylinskii et al., 2016).

7.5.1.1 Area under the ROC curve

Our key metrics for the model evaluation are the ones based on the receiver operating characteristic (ROC) which is the most widely used method to evaluate and compare saliency models. There is a measure strongly related to ROC called the area under the curve (AUC) which is considered in our evaluation, as well.

Considering the saliency model as a binary classifier at different thresholds of saliency values, the model can be evaluated by the ROC metrics. Typical binary classifier characteristics are true-positive (R_{TP}) and false-positive rates (R_{FP}) summing up its performance. These rates are defined as:

$$R_{FP} = \frac{F_P}{F_P + T_N},$$

$$R_{TP} = \frac{T_P}{T_P + F_N}$$

where T_P (true-positive) denotes salient and fixated pixel, T_N (true-negative) denotes non-salient and non-fixated pixel, F_P (false-positive) denotes salient and non-fixated pixel, F_N (false-negative) denotes non-salient and fixated pixel. The ROC curve means the trade-off between R_{FP} and R_{TP} at various thresholds. This trade-off can be easily visualized as on Figure 7-19. The higher R_{TP} for every R_{FP} value is, the better the model scores.

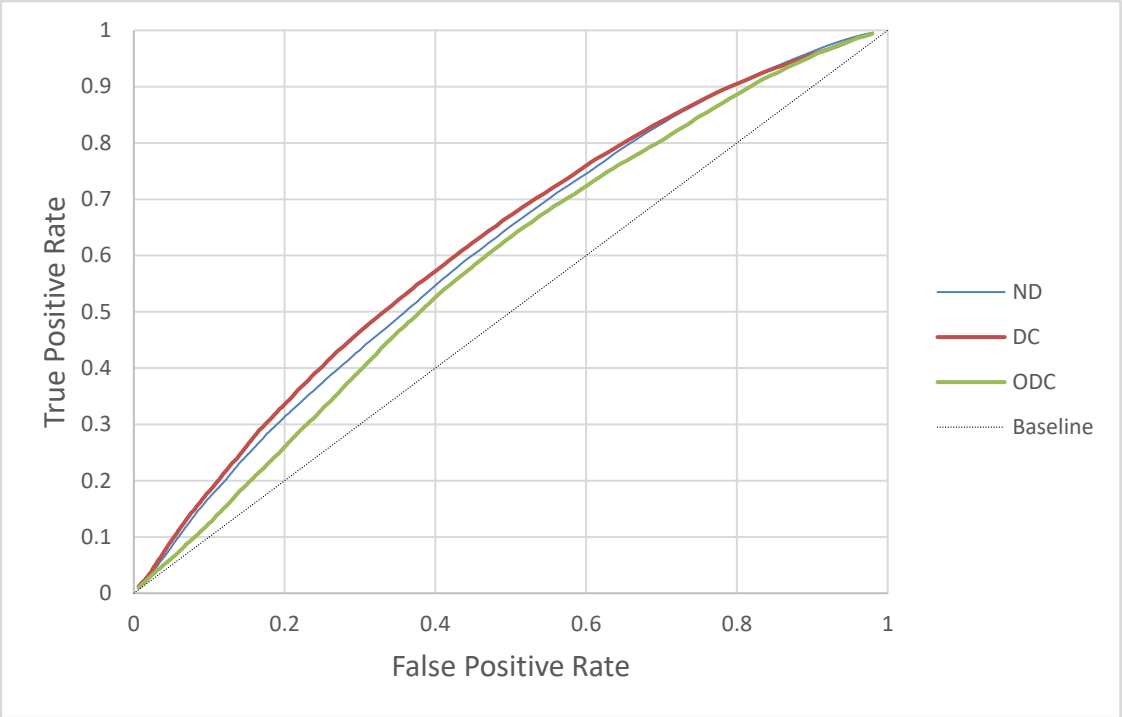


Figure 7-19: ROC curves from the evaluation of proposed model in the master thesis of Olešová (Olešová, 2016). Each ROC curve represents one saliency model performance. ND means ”no depth” module, DC means depth-contrast module and ODC means depth contrast module enhanced by the depth-saliency function proposed by Olešová.

The AUC metric is the area under the ROC curve which better summarizes the mentioned R_{TP} - R_{FP} trade-off and describes the model score as a single value. It may be computed as an integral of the trade-off function in the range $\langle 0;1 \rangle$ or simply as an approximation of the area under the curve. The approximation sums up the areas under the discrete tuples of the trade-off values and is accurate enough for our evaluation purposes.

An ideal saliency model, perfectly predicting the saliency values, should have the AUC value equal to one. Completely random saliency model, not able to distinguish saliency better than the random distribution of saliency values, has the AUC value equal to one half (baseline on Figure 7-19). Practical meaning of the AUC metric may be in praxis interpreted as a task for the model when given two locations the model has to choose the one corresponding to fixation (Bylinskii et al., 2016).

7.5.1.2 Normalized scan-path saliency

Different metric for more complex model evaluation is normalized scan-path saliency metric (NSS). With the help of a simple interpretation of the metric by Bylinskii, we can say that NSS is the average normalized saliency at fixation locations:

$$\mathrm{NSS}(\mathrm{P}, \mathrm{QB}) = \frac{1}{N} \sum_{i} \overline{P}_i \times Q_i^B$$

where P is the saliency map, Q^B is the ground-truth map of fixation locations and N is the number of fixated pixels (Bylinskii, 2016).

7.5.2 Evaluation results

We evaluated our visual attention models and the comparative state-of-the-art ones described in the Subchapters 7.4 and 7.5 using the metrics described in the Subchapter 7.5.1. The ROC curves of the visual attention model results on the complete NUS-3D saliency dataset are visualized on Figure 7-20. We can see that the ROC curves, for models proposed by Olešová, differ from the ones on Figure 7-19 in the previous subchapter. The difference is caused by our decision to use the official published AUC-Judd ROC metrics (Judd et al., 2012), rather than the AUC implementation by Olešová (Olešová, 2016). The AUC-Judd scores and the NSS metric comparisons are visualized side-by-side on Figure 7-21.

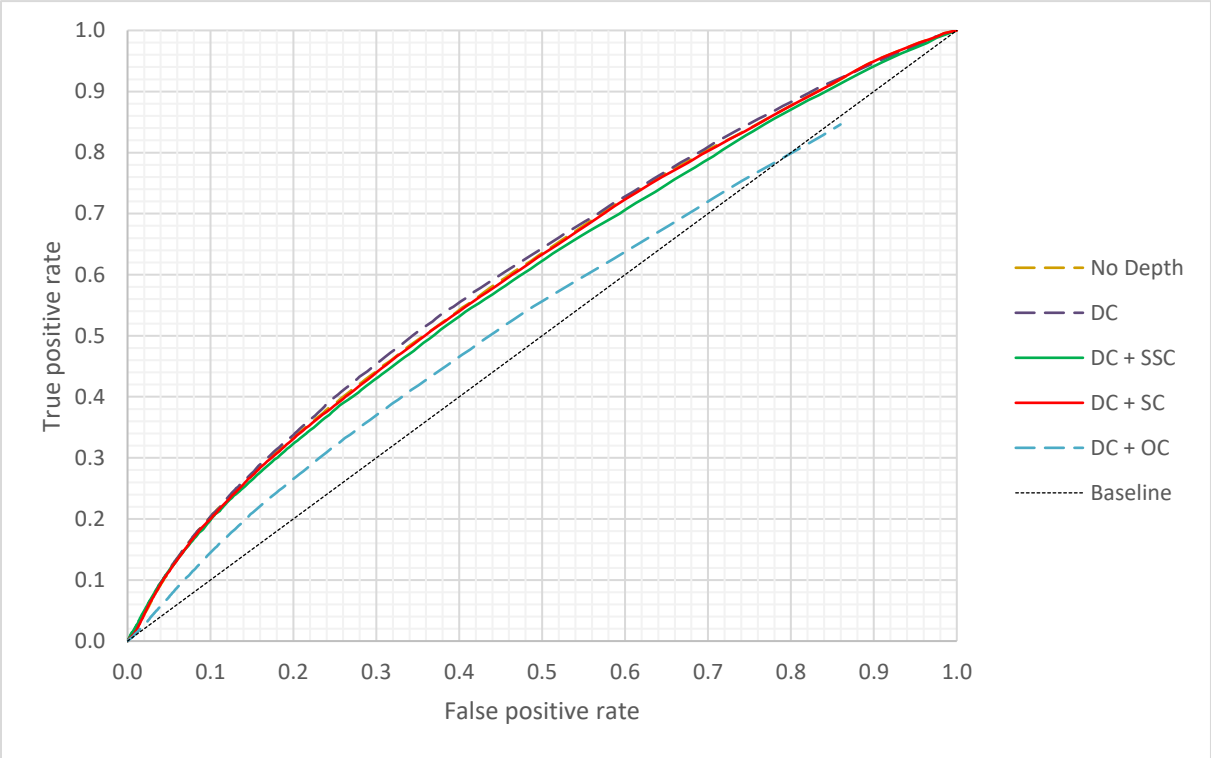


Figure 7-20 ROC curves of the five evaluated visual attention models of which two are ours, implementing the proposed depth saliency coefficient (DC+SCC and DC+CS). Three other models evaluated for comparison purposes (Olešová, 2016) are long-dashed and baseline of pure random model is black-dotted.

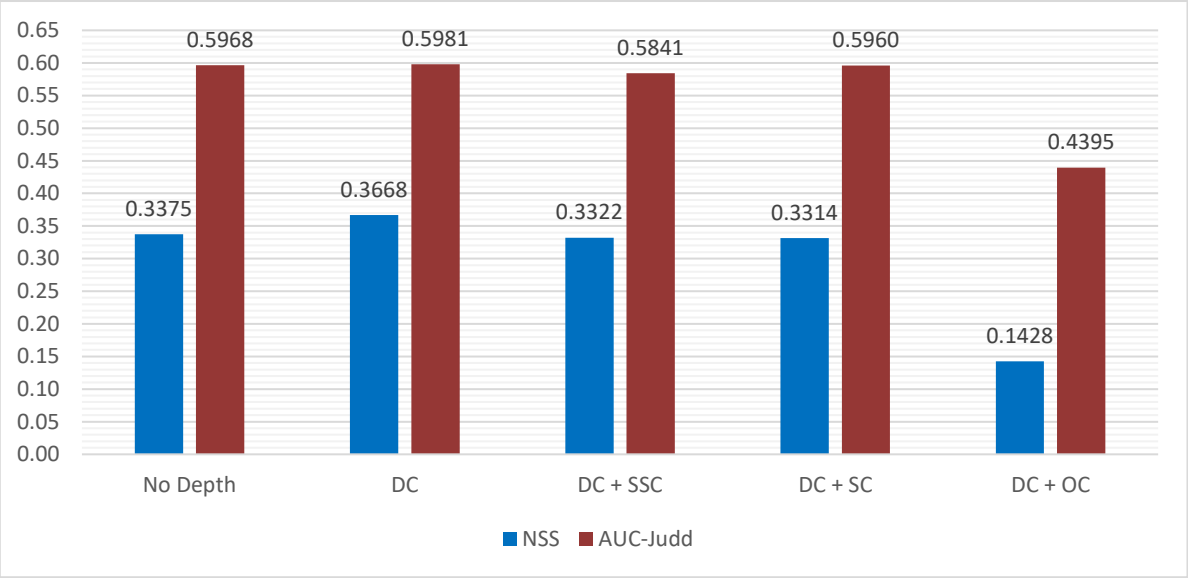


Figure 7-21 Normalized saliency scan-path and AUC-Judd metrics (Judd et al., 2012) visualized for each of the evaluated visual attention models. Two of the models (DC+SSC, DC+SC) implement our proposed depth saliency coefficient, others are listed for comparison.

We can see from the provided ROC curve diagram on Figure 7-20 that nearly all the evaluated models have better results than the baseline, meaning a model returning fully randomized distribution. The worst results encountered were the one of the depth contrast saliency model enhanced with the depth influence function proposed by Olešová (DC+OC). Resulting curve of the model is even below the curve of saliency model where the depth information is not incorporated at all. This means that the proposed function by Olešová was not accurate enough to provide good results on the evaluation dataset. We have to emphasize a significant step forward in the depth-saliency function modelling as our proposed models enhanced with the depth saliency coefficient both outperforms the model incorporating depth influence function proposed by Olešová.

Another important comparison milestone for our evaluation is the mentioned model where the depth information is not incorporated at all. Our goal was to outperform this one and not to score worse because of our claim that depth plays significant role as an aspect of human visual attention. The goal was nearly achieved by one of our models – the depth contrast model enhanced with weighting each saliency map value by depth saliency coefficient using depth information at each point of the image (DC+SC). The ours DC+SCC model scored worse due to considering mean superpixel depth during the saliency coefficient computation which may not be accurate in most cases.

Looking at the AUC-Judd score visualization on Figure 7-21, we can say that the performance of the models was nearly equivalent. However, the DC+SC provided worse results as the basis for this model before the depth enhancement- the depth contrast one (DC). The results are worse by only approximately 0.02 of AUC-Judd score which is a very little difference that can be considered as a tie of the results. This implies that we were not able to enhance performance where the depth information was already implemented in a relative way. We, however did not make the results worse. By looking at the histograms of AUC-Judd scores evaluated on the dataset (Figures 7-22 and 7-23), we can say that our proposed model has nearly a normal distribution of scores, unlike the DC model which has the distribution more concentrated around the mean value of 0.5981 and has some outliers with very low scores. Therefore, we state that we successfully normalized the DC model performance on the dataset with our enhancement.

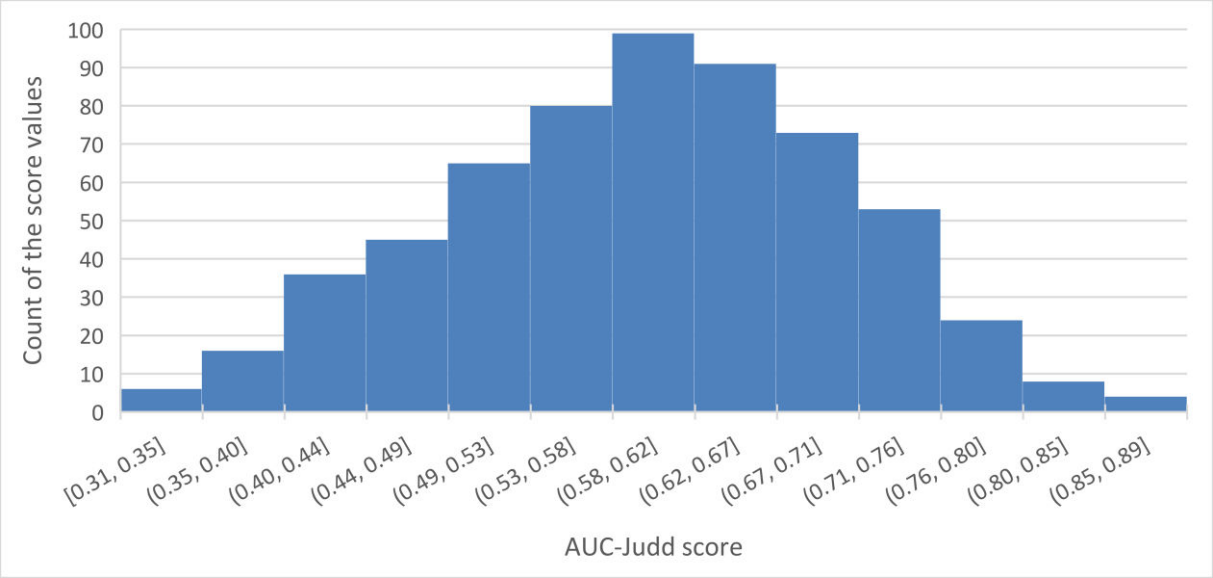


Figure 7-22 Histogram of AUC-Judd scores by our proposed DC+SC model on the NUS-3D dataset. The histogram shows nearly normal Gaussian distribution of the scores.

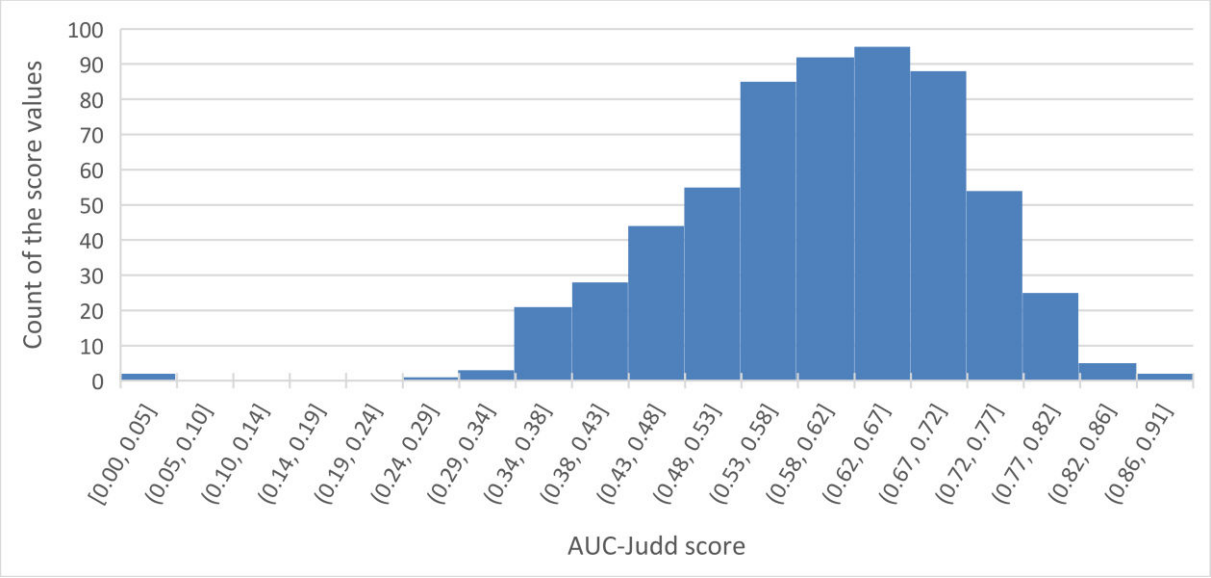


Figure 7-23 Histogram of AUC-Judd scores by compared DC model (proposed by Olešová) on the NUS-3D dataset. The histogram shows unequally distributed scores with some outliers.

We analysed the evaluation outputs of our best performing model- the DC+SC one. By visual analysis with the help of AUC-Judd score chart for the whole dataset, we made some assumptions on the aspects that may influence the results of our model. We can see on Figure 7-24 that our model was able to perform very good on certain images (AUC-Judd score above 0.85) and was enhancing the DC model significantly. This was the case when both our basis DC model provided satisfactory results and the depth map of the scene was accurate. One can also spot that reliable results were achieved indoors with the scenery depth range approximately matching the one of our laboratory where the major experiment took place. This could be because of best results of Kinect depth maps in such conditions but also because of the depth saliency coefficient variance on the depth range of the scene.

The first possible reason- quality of the Kinect's depth map- influenced the results of our model on the dataset significantly. We consider part of the dataset as inappropriate for publishing along with the depth

map as the depth map was captured in sceneries with the depth range significantly out of the Kinect depth map range (scenery range from 0 to 50 meters in some cases; Kinect depth map range of approximately 0.4 to 8 meters with certain inaccuracies from about 4 meters[14]). The information about depth of the scene is essential source of information for our proposed depth saliency coefficient. The weak side of the NUS-3D dataset is, however, the depth information captured by Kinect and stored as grayscale images with 8-bit depth. The 8-bit resolution with no additional information about each of the depth maps (depth range in the depth map, calibration value, etc.) makes the depth values only informative, rather than exact for such a computation as our depth saliency coefficient. Unfortunately, datasets with the exact depth information captured by high quality devices are not available, nowadays. Therefore, we cannot provide the evaluation results with higher confidence than just the informative and approximative one. The example of unusable depth maps in the NUS-3D dataset are visualized on Figure 7-26.

The second plausible reason for our enhanced model scores- depth saliency coefficient variance on the depth range of the scene- is a wider research problematic that resulted in our assumption that we state as one of the conclusions in our work. The most salient depth exists, and it might be greatly influenced by the depth range of the scene. Thus, further research should be focused on the scene's depth range influence on the most salient depth and its application on the existing models to make them invariant on the scene size. The inaccurate saliency maps, partially due to the various depth ranges of the scenes, are captured on Figure 7-25.

There are a few other factors that have certain impact on the evaluation results of our enhanced visual attention models that are listed in this subchapter. One of them is the reliability of the visual attention model that we took as the basis for our enhanced model including weighting by the depth saliency coefficient. The output of the DC model is far not as good as the current state-of-the-art visual attention models and is not reliable in many cases, with scores that are often worse than the baseline (in approximately 1/3 cases as on Figure 7-23). The fact is, that depth saliency coefficient cannot improve these inaccurate results in a significant way and, thus, is not scoring better on at least 1/3 of the dataset. The example of bad scoring DC model and, hence, bad scoring ours, enhanced model is visualized on Figure 7-25. As we stated before, the DC model was chosen as the basis for our enhanced models because of the comparison purposes with the previous work we built upon. Hence, we should implement our depth saliency coefficient in currently best saliency models (these are listed in MIT saliency benchmark[15]) for more accurate and reliable evaluation and results. The best model's implementations are, however, not public in most cases.

Overall, we conclude, regarding the evaluation results, that they are informative, approximative, and serve as a comparison with the previous work we built upon and as a proof of depth saliency coefficient relevancy when applied on the existing saliency model. Our evaluation shortfall is the strong dependency of our proposed model enhancement on reliable results of the existing saliency model and on the accurate information about the depth of the scene. Moreover, we claim that there are no possibilities to evaluate the depth saliency coefficient in the environments it was proposed for (real environments), nowadays. Datasets for studying human visual attention in real environments from the egocentric perspective of view, including the depth information, are not available these days (except the novel dataset used for depth saliency coefficient proposal).

[14] Available on 21/04/2018 at: https://msdn.microsoft.com/en-us/library/hh973078.aspx#Depth_Ranges
[15] Available on 22/04/2018 at: http://saliency.mit.edu/results_mit300.html

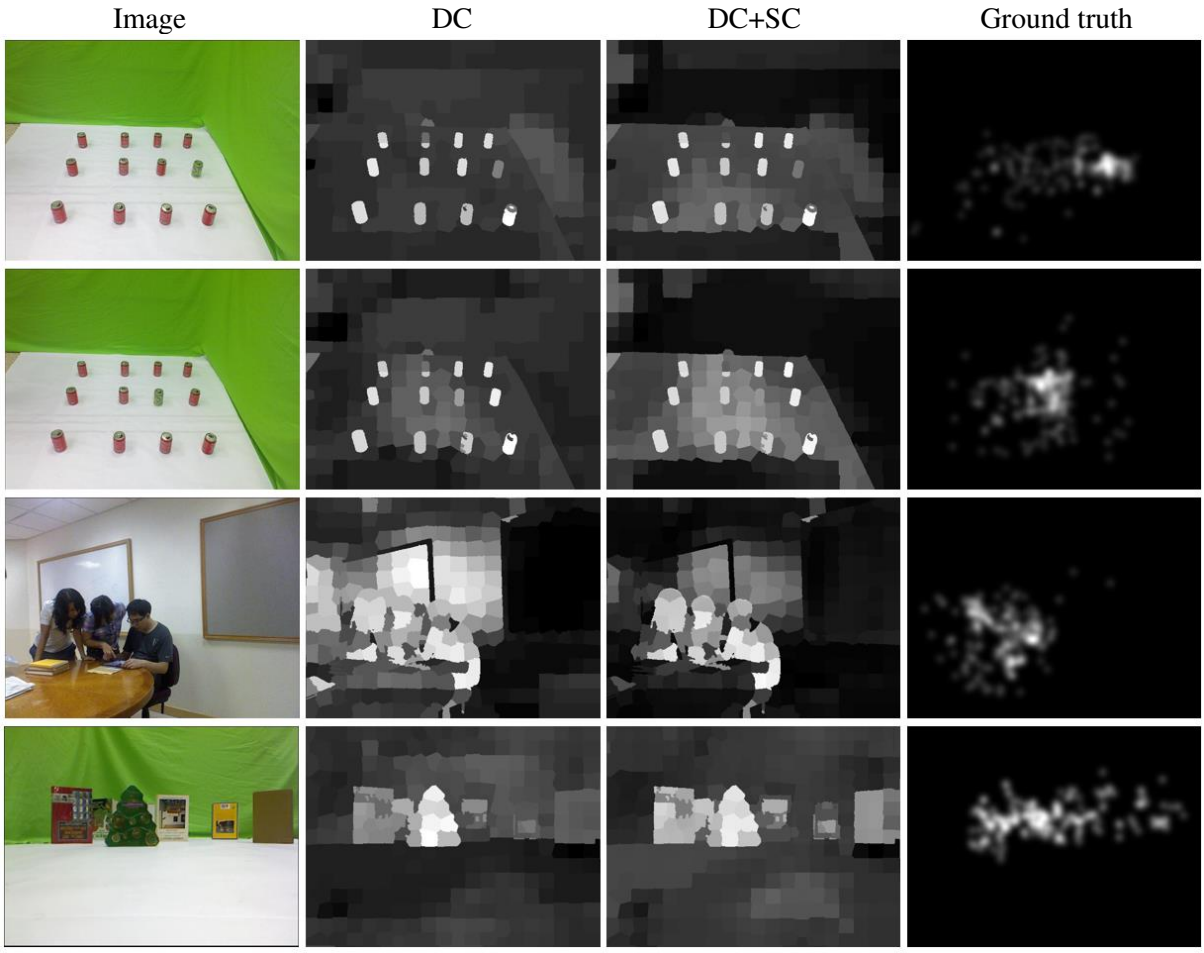


Figure 7-24 Saliency maps generated from the input image (a) by the DC model (b), DC+SC model (c) and the ground truth of human fixations on the stereoscopic 3-D image (d). The depth saliency coefficient improved the accuracy of the DC model significantly. Even better results may be achieved if we would build upon a reliable saliency module and would have accurate depth information.

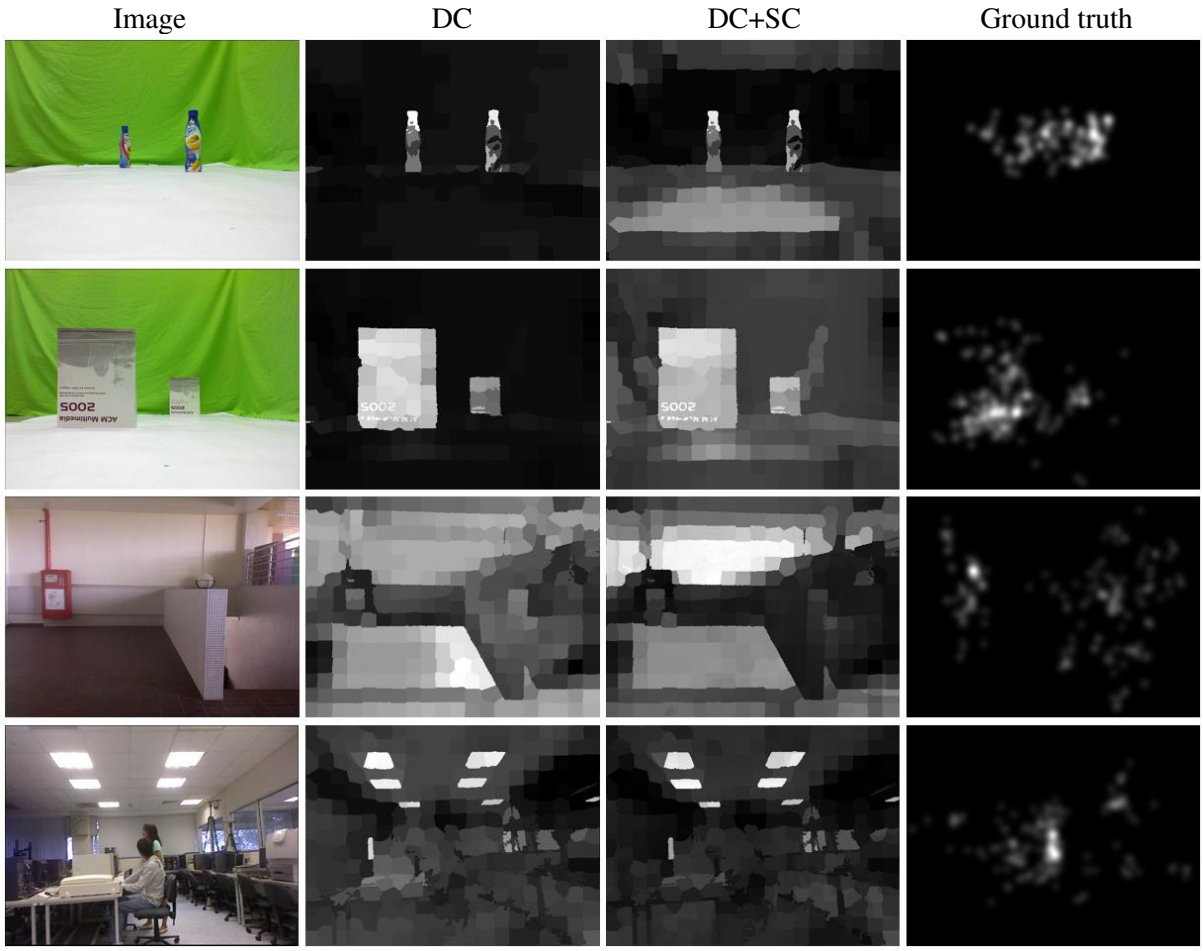


Figure 7-25 Saliency maps generated from the input image (a) by the DC model (b), DC+SC model (c) and the ground-truth of human fixations on the stereoscopic 3-D image (d). We can see that the depth saliency coefficient does not perform well in the scenes with different depth range than the one of the laboratory where our research was held (the first and second image). Influence of the depth range of the scene on the depth saliency coefficient is an open research topic. Moreover, our enhancement of the DC module does not perform well if the results from the base model are not accurate enough (the third and fourth image).

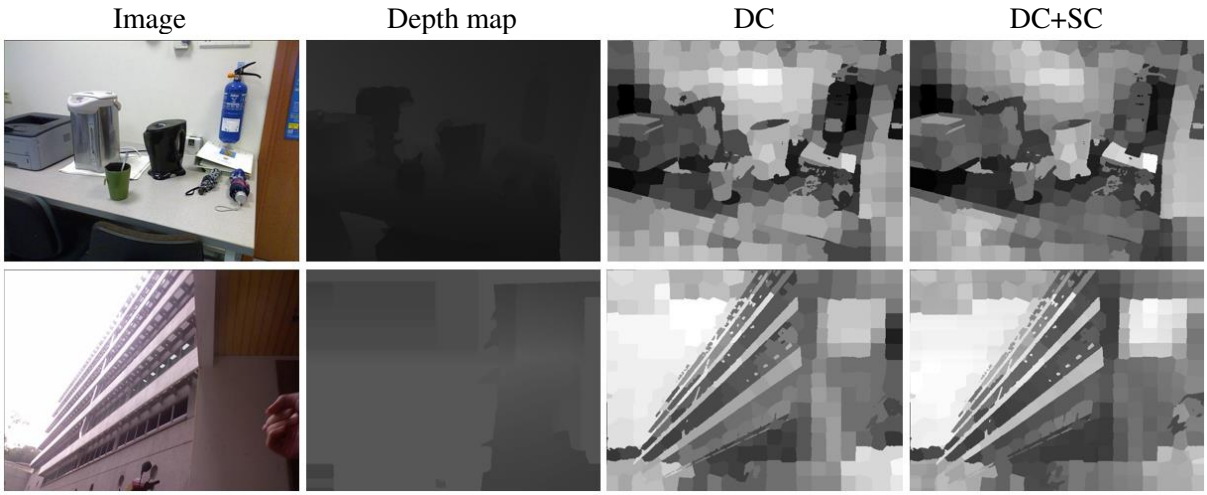


Figure 7-26 Saliency maps generated from the input image (a) and the depth map (b) by the DC model (c), DC+SC model (d). We can see that the depth saliency coefficient is directly dependent on the accurate depth information of the scene. However, the depth information for some images in the NUS-3D dataset is unusable and mispresent the results of our proposed module.

8 Conclusions

We provided the analysis on the theory base related to human visual attention in the analytical part of this thesis. We focused the analysis on visual attention modelling and its techniques and approaches to study the human visual attention from the technical point of view. We thoroughly studied numerous state-of-the-art research papers related to the visual attention modelling. We chose a few strongly related papers to this thesis and described them in the state-of-the-art part of the thesis which we further built upon. We emphasized novelties and trends in the visual attention modelling and pointed out open topics related to the visual attention which are under active research, nowadays. Based on the knowledge obtained from the analytical part of our work, we summarized the knowledge we built upon to set goals and state assumptions for our research.

We proposed a novel method to conduct user studies for the research of visual attention in real world environments and introduced a new approach to study visual attention on a real scene in a laboratory with possibility of dynamic changes at the scene. We proposed our solution as an automated and adjustable methods including the user study methodology and the user study setup proposal. Our proposal introduces repeatable, scalable and versatile user studies with human visual attention in real environments, suitable not only for this thesis but also for future research of various aspects affecting human visual attention in everyday life. To achieve such a goal, we introduced several own procedures, algorithms and modifications of the existing ones using knowledge and principles of computer vision.

Our research method proposal, including software modules necessary for supporting the user studies, were implemented, tested during numerous experiments and used for extensive user studies. We held the major experiments following our proposed method in a laboratory with 37 participants to create a novel dataset for further research of the human visual attention in real environments from the egocentric perspective of view. Using the created dataset, we focused our research on the aspect greatly influencing visual attention from the perspective of the observer in real environments- depth of the scene.

We summed up the results of our research by statistical evaluation of the first fixations of the observers after occurrence of dynamic changes at the scene. Results of the evaluation are visualized and discussed. Our research supports the claim that depth plays significant role as the aspect greatly influencing human visual attention. We claim that the influence can be approximated as a so-called depth saliency coefficient applicable on existing models of human visual attention. Moreover, we claim that there exists the novel phenomenon called the most salient depth- range of the distances from the observer where objects have the highest saliency for the observer. The saliency of the objects then decreases significantly with their distance from the most salient depth.

The results of the depth saliency coefficient applied on the state-of-the-art saliency model, considering the saliency from the camera perspective, are summed up in the last chapter of this thesis. We claim that we were able to define the absolute depth influence on human visual attention better than any of the known state-of-the-art models. The best scoring among our proposed models was the depth contrast model enhanced with the depth saliency coefficient (incorporating originally only the relative depth of the captured image). The results of our models were satisfying and could have been even better if the depth saliency coefficient was applied on more reliable visual attention model and evaluated on the dataset with high quality information about the depth of the scene, or on the dataset studying human visual attention in real environments. Unfortunately, this kind of dataset is not available, nowadays, and the one created by us is a novelty in the field of the visual attention modelling.

We assume that application of the knowledge in this thesis on the existing models of visual attention will produce more accurate models of human visual attention and better reflect saliency of the objects in real environments. However, further research using our proposed method is necessary to prove our assumption. The proposed depth saliency coefficient should be applied on a reliable saliency model and evaluated on the dataset of a higher quality, mainly speaking about the depth information and ground truth data better reflecting observations of the scene in reality. Moreover, further extensive user studies following our research method proposal could lead to novel knowledge about human visual attention and may denote how to better incorporate knowledge from such a research into the existing saliency models. In the end, we claim that the most salient depth exists at a scene and that it might be greatly influenced by the depth range of the scene. Thus, future research may be focused on the scene's depth range influence on the most salient depth and its application on the existing models to make them invariant on the scene size.

9 References

BANKS, S. J., et al.: Amygdala–frontal connectivity during emotion regulation. Social cognitive and affective neuroscience, 2007. pp. 303-312.

BARRON, J. L., et al.: Performance of optical flow techniques. In: Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on. IEEE, 1992. pp. 236-242.

BEAR, M. F., CONNORS, B. W., PARADISO, M. A.: Neuroscience. Lippincott Williams & Wilkins, 2007.

BIRKUS, R.: Image segmentation using graphics processing unit. Bachelor thesis. Bratislava: FIIT STU, 2015. 40 p.

BORJI, A., ITTI, L.: State-of-the-art in visual attention modeling. IEEE transactions on pattern analysis and machine intelligence, 2013. pp. 185-207.

BOUJUT, H., BENOIS-PINEAU, J., MEGRET, R.: Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion. In: European Conference on Computer Vision. Springer Berlin Heidelberg, 2012. pp. 436-445.

BRUHN, A., WEICKERT, J., SCHNÖRR, C.: Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. International journal of computer vision, 2005. pp. 211-231.

BUSO, V., GONZÁLEZ-DÍAZ, I., BENOIS-PINEAU, J.: Goal-oriented top-down probabilistic visual attention model for recognition of manipulated objects in egocentric videos. Signal Processing: Image Communication, 2015. pp. 418-431.

BYLINSKII, Zoya, et al.: What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1604.03605, 2016.

CONGYAN, L., et al.: Depth Matters: Influence of Depth Cues on Visual Saliency. ECCV (2) 2012. pp. 101-115

DENIL, M., et al.: Learning where to attend with deep architectures for image tracking. Neural computation, 2012. pp. 2151-2184.

FARNEBÄCK, G.: Two-frame motion estimation based on polynomial expansion. Image analysis, 2003. pp. 363-370.

FATHI, A., REN, X., REHG, J. M.: Learning to recognize objects in egocentric activities. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On. IEEE, 2011. pp. 3281-3288.

FISTER JR, I., et al.: A brief review of nature-inspired algorithms for optimization. arXiv preprint arXiv:1307.4186, 2013.

GARRIDO-JURADO, S., et al.: Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognition, 2014, 47.6: 2280-2292.

GOLDSTEIN, E. B.: The Blackwell Handbook of Sensation and Perception. John Wiley & Sons, 2008.

HAFED, Z. M., CLARK, J. J.: Microsaccades as an overt measure of covert attention shifts. Vision research, 2002. pp. 2533-2545.

HAJCAK, G., et al.: The dynamic allocation of attention to emotion: simultaneous and independent evidence from the late positive potential and steady state visual evoked potentials. Biological Psychology, 2013. pp. 447-455.

HAREL, J., KOCH, C., PERONA, P.: Graph-based visual saliency. In: Advances in neural information processing systems. 2007. pp. 545-552.

HAO, J., et al.: Visual attention deficits in Alzheimer's disease: an fMRI study. Neuroscience letters, 2005. pp.18-23.

HORN, B. K. P., SCHUNCK, B. G.: Determining optical flow. Artificial intelligence, 1981. pp. 185-203.

ITTI, L., KOCH, C., NIEBUR, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence, 1998. pp. 1254-1259.

ITTI, L.: Models of bottom-up and top-down visual attention. PhD Thesis. California Institute of Technology, 2000.

ITTI, L., KOCH, C.: Computational modelling of visual attention. Nature reviews neuroscience, 2001. pp. 194-203.

ITTI, L., DHAVALE, N., PIGHIN, F.: Realistic avatar eye and head animation using a neurobiological model of visual attention. In: Optical science and technology, SPIE's 48th annual meeting. International Society for Optics and Photonics, 2004. pp. 64-78.

ITTI, L.: Automatic foveation for video compression using a neurobiological model of visual attention. IEEE Transactions on Image Processing, 2004. pp. 1304-1318.

JIANG, M., et al.: Salicon: Saliency in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. pp. 1072-1080.

JUDD, T., DURAND, F., TORRALBA, A.: A benchmark of computational models of saliency to predict human fixations. 2012.

KARTHIKEYAN, S.; JAGADESCH, V.; SHENOY, R.; ECKSTEINZ, M.; MANJUNATAH, B. S. From Where and How to What We See, 2013 IEEE International Conference on Computer Vision, Sydney, NSW, 2013. pp. 625-632. doi: 10.1109/ICCV.2013.83

KHOSHELHAM, K.; ELBERINK, S. O. Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors, 2012, 12.2: 1437-1454.

KIENZLE, W., et al.: Center-surround patterns emerge as optimal predictors for human saccade targets. Journal of vision, 2009. pp. 7-7.

KLEIN, D.A.; FRINTROP, S.: Center-surround divergence of feature statistics for salient object detection. In Computer Vision (ICCV), 2011 IEEE International Conference IEEE. pp. 2214-2219

KOCH, C., ULLMAN, S.: Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of intelligence. Springer Netherlands, 1987. pp. 115-141.

KRUTHIVENTI, S.S., AYUSH, K., BABU, R. V.: Deepfix: A fully convolutional neural network for predicting human eye fixations. arXiv preprint arXiv:1510.02927, 2015.

KÜMMERER, M., WALLIS, T., BETHGE, M.: DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:1610.01563, 2016.

LANG, C., et al.: Depth Matters: Influence of Depth Cues on Visual Saliency. ECCV (2) 2012. pp. 101-115

LE MEUR, O., LE CALLET, P., BARBA, D.: Predicting visual fixations on video based on low-level visual features. Vision research, 2007. pp. 2483-2498.

LI, J., et al.: Probabilistic multi-task learning for visual saliency estimation in video. International journal of computer vision, 2010. pp. 150-165.

LIU, H., et al.: A generic virtual content insertion system based on visual attention analysis. In: Proceedings of the 16th ACM international conference on Multimedia. ACM, 2008. pp. 379-388.

LUCAS, B. D., et al.: An iterative image registration technique with an application to stereo vision. 1981.

MARCHESOTTI, L., CIFARELLI, C., CSURKA, G.: 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.

MATSUO, K., 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.

MERKER, B.: Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. Behavioral and brain sciences, 2007. pp. 63-81.

MIRJALILI, S.: Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. Knowledge-Based Systems, 2015. pp. 228-249.

MIRJALILI, S., MIRJALILI, S. M., HATAMLOU, A.: Multi-verse optimizer: a nature-inspired algorithm for global optimization. Neural Computing and Applications, 2016. pp. 495-513.

MITRI, S., et al.: Robust object detection at regions of interest with an application in ball recognition. In: Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on. IEEE, 2005. pp. 125-130.

MNIH, V., et al.: Recurrent models of visual attention. In: Advances in neural information processing systems. 2014. pp. 2204-2212.

MURRAY, N., et al.: Saliency estimation using a non-parametric low-level vision model. In: Computer vision and pattern recognition (cvpr), 2011 IEEE conference on. IEEE, 2011. pp. 433-440.

NAGAI, Y., MUHL, C., ROHLFING, K. J.: Toward designing a robot that learns actions from parental demonstrations. In: Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on. IEEE, 2008. pp. 3545-3550.

OLEŠOVÁ, V.: Generating a Saliency Map using Depth Information. Master thesis. Bratislava: FIIT STU, 2016. 65 p.

OLSEN, A.: The tobii i-vt fixation filter. Tobii Technology, 2012. 21 p.

PHAN, K. L., et al.: Neural substrates for voluntary suppression of negative affect: a functional magnetic resonance imaging study. Biological psychiatry, 2005. pp. 210-219.

PIERROT-DESEILLIGNY, C., MILEA, D., MÜRI, R. M.: Eye movement control by the cerebral cortex. Current opinion in neurology, 2004. pp. 17-25.

POLATSEK, P.: Spatiotemporal Saliency Model of Human Attention in Video Sequences. Master thesis. Bratislava: FIIT STU, 2015. 78 p.

REN, X., GU, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference. pp. 3137-3144.

RENSINK, R. A., O'REGAN, J. K., CLARK, J. J.: To see or not to see: The need for attention to perceive changes in scenes. Psychological science, 1997. pp. 368-373.

RENSINK, R. A.: Change detection. Annual review of psychology, 2002. pp. 245-277.

ROBERTS, K. L., et al.: Visual search in depth: The neural correlates of segmenting a display into relevant and irrelevant three-dimensional regions. NeuroImage, 2015. pp. 298-305.

SCHÖLKOPF, B., SMOLA, A. J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002.

SIMONS, D. J.; CHABRIS, C. F.: Gorillas in our midst: Sustained inattentional blindness for dynamic events. Perception, 1999. pp. 1059-1074.

SIMONYAN, K., ZISSERMAN, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

SPILLMANN, L., WERNER, J. S.: Visual perception: The neurophysiological foundations. Elsevier, 2012.

SRIPADA, C., et al.: Volitional regulation of emotions produces distributed alterations in connectivity between visual, attention control, and default networks. Neuroimage, 2014. pp.110-121.

SUZUKI, S., et al. Topological structural analysis of digitized binary images by border following. Computer vision, graphics, and image processing, 1985, 30.1: 32-46.

TREISMAN, A. M., GELADE, G.: A feature-integration theory of attention. Cognitive psychology, 1980. pp. 97-136.

TREUE, S.: Neural correlates of attention in primate visual cortex. Trends in neurosciences, 2001. pp. 295-300.

UNGERLEIDER, S. K., LESLIE, G.: Mechanisms of visual attention in the human cortex. Annual review of neuroscience, 2000. pp. 315-341.

VERRI, A., POGGIO, T.: Motion field and optical flow: Qualitative properties. IEEE Transactions on pattern analysis and machine intelligence, 1989. pp. 490-498.

VIG, E., DORR, M., COX, D.: 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.

WANG, J., DA SILVA, M.P., LE CALLET, P., RICORDEL, V.: Computational model of stereoscopic 3D visual saliency. IEEE Transactions on Image Processing, 22(6), 2013. pp. 2151-2165.

XU, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. 2015. p. 2048-2057.

YAMADA, K., et al.: Attention prediction in egocentric video using motion and visual saliency. Advances in image and video technology, 2012. pp. 277-288