Depth in the visual attention modelling from the egocentric perspective of view

Miroslav Laco*
Supervised by: Wanda Benesova†

Institute of Computer Engineering and Applied Informatics
Slovak University of Technology
Bratislava / Slovakia

Abstract

A research goal of computer vision scientists in the field of human visual perception is to determine and predict visual attention by the creation of various models of visual attention. An extensive research has been held in the field of visual attention modelling throughout the past years, aiming to create a complex visual attention model as close to ground-truth as possible.

In this paper, we mention the benefits of user studies on the human visual attention in real world environments from the egocentric perspective of view and we introduce a novel and complete method proposal for such user studies in a laboratory. We make use of specific hardware equipment (eye-tracking glasses, Kinect and LCD projectors) and introduce our own algorithms and procedures based on computer vision theory in our method proposal. The implementation of a novel method for user studies resulted in a small novel dataset created during the first experiments.

One of the benefits of the proposed novel method is the possibility to study aspects affecting visual attention that were not possible to study before. Based on the previous research in the field, we decided to conduct a research of depth influence (distance between the observer and the observed object) on human visual attention using created dataset. We claim that the depth of the scene plays significant role as an aspect of human visual attention in real world environments.

Keywords: Human visual attention, Egocentric perspective of view, Visual attention modelling, Saliency

1 Introduction

Human visual perception make up to 80% of all perceived sensual information entering our brain [1]. Despite of the fact, the amount of information getting to our central nervous system through the eyes is much higher than the processing capacity of our brain. This is a reason why the visual perception system of humans consists of a selection mechanism applied on the perceived stimuli and a notion of relevance [1]. Selectivity is a part of complex process of visual attention which helps us to decide where to fix our eyes and which visual stimuli to process first by determining salient regions [4]. Scientific research related to visual attention is interdisciplinary and involves work of psychologists, neurobiologists and computer vision scientists. The research aim of computer vision scientists is to determine and predict visual attention by models of visual attention [1].

Visual attention models consider stimuli that affect human visual attention- the so-called aspects of human visual attention. These aspects can be divided into two main groups: aspects related to properties of a scene itself (colour, size, orientation of objects, etc.) and aspects related to subjective state of the observer (previous knowledge, memory factors, visual task, etc.). Throughout the past years, the research in the field and the experiments for user studies of visual attention has been making use of standard displaying devices (e.g. projected scene on a monitor) to approximate impact of each of these aspects on human visual attention. Nowadays, modern hardware and various methods of computer vision can be used to create novel methods for a user study of visual attention in real world environments and, thus, study aspects affecting visual attention that were not possible to study before.

In this paper, we introduce a novel method for user studies focused on research of the human visual attention in a real-world environments. This means that we ignore specifics of visual attention modelling from the camera perspective, which are well known from numerous research papers (some of them mentioned in section 2), and introduce new perspective (the egocentric one) for the research of human visual attention. The ultimate goal of our novel research approach, possibly in the future work, is to implement our findings from the egocentric perspective in an existing saliency model which is based on the previous visual attention research from the perspective of the camera. This is the biggest asset in the field of the human visual attention research that we present in this paper. The novel approach leads to some novelties in the way we look at the saliency of the objects in the scene. The common

*miroslav.laco@gmail.com
†benesova@it.stuba.sk

saliency model from the camera perspective looks at the static objects in an image through their planar properties-size, colour, orientation, contrast, etc. Saliency of the object is computed by the model based on these properties. However, taking into account real environments and the egocentric perspective (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 (meaning they itself have the same saliency). The saliency of the static objects perceived by the individual observer from his egocentric perspective is then influenced by the objects position in the scene and the internal state of the observer [9, 11, 7]. This means that the position of the objects in the scene is the most significant bottom-up aspect affecting saliency of the static objects in the scene from the observer's egocentric perspective of view. The internal state of the observer is not taken into account in this paper.

Moreover, we introduce a method for application of the novel findings from the research of human visual attention in the real environments into the existing saliency models based on the camera perspective. When speaking about the position of the objects in the scene and its influence on their saliency, the influence of 2-D position of the objects on the scene on their saliency can be sufficiently described by the existing saliency models (Borji et al. [1]), nowadays. Therefore, we will focus our research mainly on the third dimension of the object positions in the scene and its impact on their saliency. Moreover, we claim that the saliency of the object itself on the scene can be determined with high accuracy by the existing saliency models, too. Then, the aspect of a position of the objects in the scene (in our case the third dimension- depth of the scene) can be applied on the determined saliency of the objects by the conventional saliency models as a weighting function of a 3-D position of the object in the scene. We study this phenomenon on the dataset created during the first user studies, following our novel proposed research method and regarding the novel research approach described in previous paragraph, and present first results of our novel approach in this paper.

Related research papers from the state-of-the-art of visual attention modelling, we build upon when formulating our claims, are outlined in the Section 2. Our proposed method is described more in detail in the Section 3. Further details about methods and algorithms we introduce to achieve goals of the proposed method are described in the Section 4. The results of our research are evaluated and discussed in the Section 5. Furthermore, our contribution to the research and modelling of visual attention and the possible future work is summarized in the Section 6.

2 Related work

The majority of research methods conducting user studies to study human visual attention has been making use of conventional displaying devices (monitor, LCD projector, virtual reality displays) for displaying some scene or images (either 2-D or 3-D) to an observer. Captured information about visual attention of the observers, while looking at the displayed scenes, were used for creating novel models of visual attention or enhancing the existing ones. In the previous years, it has been proved that the third dimension of a scene (its depth) plays significant role as an aspect of human visual attention [1].

Similar research methodology, as described in the previous paragraph, was proposed by Wang et al. [11]. The idea of the proposed method was to display a novel dataset of stereoscopic 3-D images of natural scenes to the observers during their user studies except of the previous, conventional, 2-D ones. They were one of the first pioneers studying third dimension (depth of the scene) in relation to the visual attention. Similar method was chosen in the research of Lang et al. [6] where the correlation of visual attention between observed 2-D and 3-D images was studied. The NUS-3D saliency dataset created during their user studies is a significant benefit of their work where saliency in 3-D scenes from the camera perspective can be studied. We can compare results of our approach to the ground-truth of this dataset by the end of our research. Both research works were studying the bottom-up manners of visual perception assuming that depth aspect influence is related to the preattentive stage of visual perception. We build upon this knowledge and decided to study depth influence on the objects saliency from the bottom-up egocentric perspective, omitting the internal state of the observer (top-down aspects).

One of the most recent research articles discussing depth as an aspect of visual attention was held by Roberts et al. [9].It is claimed in the conclusion that a depth plays significant role in both preattentive stage and attentive stage of visual perception and, therefore, influences visual attention in bottom-up and top-down manner. This was proved by the user studies including search tasks on projected scene (visual search on the scene occurs during the attentive stage of a visual perception). Roberts et al. conclude that participants of the user study responded significantly more quickly and accurately when relative depth information about searched object on the scene was known rather than unknown. Moreover, MRI scanning of the observer's brain activities was conducted during the search tasks on the scene. It is discussed that the knowledge about the search target depth in the scene influences activity in depth-sensitive parietal regions but not in depth-sensitive visual regions of our brain. We include this knowledge to the methodology of our user studies where the influence of depth visual perception in the attentive stage is suppressed by appropriate adjustment of the user studies setup.

Furthermore, we build upon knowledge that a depth of the scene is significant aspect affecting the human visual attention but is still not thoroughly studied. Depth of the scene is a natural property of a real-world environment. Therefore, we claim that a novel research method for conducting user studies on visual attention in real world en-

vironment may provide accurate results better mirroring visual attention in reality than results from the previous works. This idea was discussed in the conclusion by Olesova [7] where a claim was made that the estimation of depth influence on the human visual attention may be approximal to a polynomial function:

$$y = 2.5x10^3 - 0.00017x^2 + 0.03x - 0.75,$$

for $x \in < 50; 350 >$ where x is the depth of the scene in centimeters and y is the depth saliency coefficient.

3 Proposed method

We propose a novel method to conduct experiments for the user studies on a visual attention in real world environments, building upon the previous work stated in Section 2. 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. 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 human visual attention.

Real environment is simulated in a laboratory which should be simplistic 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 unchanged during the projection. Therefore, projector calibration 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 in Section 4.1. The experiment is static, without a motion influence on visual attention, therefore an observer is required to stand in a specific place during the user study. Proposed laboratory setup creating real scene is illustrated in Figure 1.

Visual attention data of the observer are collected during the user study by a glass eye-tracker (more in Section 4.2). The observer is instructed to look freely at the scene while changes of the projected content on the scene are being made by the projection handling software module (Section 4.1). By free viewing task 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. Caption from the experiment setup is displayed on the Figure 2.

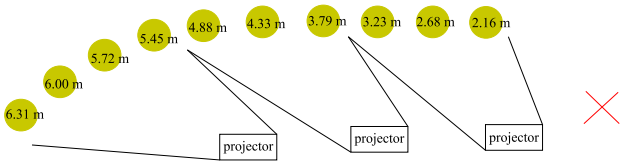

Figure 1: Schema of an experimental setup in a laboratory. A participant stands at the red cross, yellow circles represents hanging polystyrene balls (our ROIs) with distance information inside of them. We used three projectors to cover all ROIs by the projected content.

Figure 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 PC with running data acquisition module.

First dataset created by conducting the user studies following our proposed method consists of:

- the egocentric video from the observer's perspective during user study,
- the information about observer's fixations on the scene (points of concentration of visual attention in the egocentric video) in separate file,
- the information about changes on the scene and their timing which can be synchronized with egocentric video.

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 propose our own projection sequence for the research of depth influence on human visual attention. The main specificity of our projection sequence is in changing of the same projected content on two ROIs on the scene at the same time (the action is further referenced as concurrent change on ROIs). No content is projected on the ROIs at the beginning. Subsequently, the same content is projected on 2 ROIs simultaneously in different depths for duration of 1000 milliseconds with 400 milliseconds fade-in and fade-out effect. There are generated complete $\binom{10}{2}$

combinations of these concurrent changes on the scene. 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. Three types of concurrent changes on the scene are generated in random order:

1. projection of plain white colour on two ROIs on the scene (for simplicity to eliminate the influence of various aspects affecting visual attention, i.e. different colour, shape, orientation, etc.),

2. projection of slightly different colour tint on two ROIs (yellow colour *RGB(100,100,0)* projected on all ROIs and dark yellow *RGB(60,60,0)* projected as concurrent change) as suggested by Olesova [7],

3. projection of the same face on two ROIs as human face is one of the most important top-down features to attract visual attention [12].

4 Experimental setup overview

Algorithms and methods of computer vision are used to achieve goals from our method proposal and to automate significant parts of the user study setup. 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 applied principles of computer vision in the following subsections.

4.1 Projection calibration

Hanging polystyrene balls from the roof in different depths in the scene (meaning distance from a stationary observer) represent our regions of interest (ROIs) where the desired changing content is projected. We use a binary projection mask applied on a conventional LCD projectors frame buffer to project content on certain areas in the scene and nothing on other areas in the scene covered by LCD projector's projection plane. ROI background is unaffected by the projection and is not distracting participants visual attention because of the projection mask (example in Figure 3).

We are using multiple projectors across the laboratory to maintain good quality of projected content in each of the ROI at different depths. The desired state of projection on ROIs is captured in Figure 2. The creation of binary projection mask (further referenced as projector calibration) is a challenging part of the setup. We used one Kinect 2.0 device for the calibration process and proposed a calibration algorithm described in the next paragraphs.

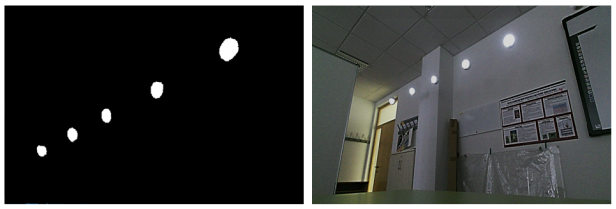

Figure 3: Projection mask after transformations (left) applied on frame buffer of projector (resulting projection on the right).

Projectors are calibrated separately and sequentially using one Kinect 2.0 device, placed in front of the calibrating device. Multiple Kinect devices are not considered as the Kinect device takes up more than 50% of a standard PC USB bus[1]. The information we collect through the Kinect device are RGB frame and normalized depth map of the scene in range $< 0;1 >$. RGB frame is used for segmentation of projection plane by its corner coordinates using either human interaction (clicks on corners) or an improved adaptive Gaussian mixture model for background subtraction (algorithm published by Zivkovic [13]). Normalized depth map from the Kinect is used for more precise segmentation of the ROIs in the scene with the help of human interaction (clicking on ROI areas) and flood-filling clicked depth level in the depth map A binary projection mask M in the depth map space is subsequently obtained as:

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

where I' is the depth map with the flood-filled regions and I is the original depth map before the flood-fill operations. The depth map space is different from the RGB image spaceof the Kinect. We have to define a geometric transformation of the binary projection mask from the depth map space to the RGB image space as the segmented projection plane is relative to the RGB image space of the Kinect. A homography matrix defining transformation of the RGB frame to the depth point-cloud space is computed for these purposes. We obtained planimetric object coordinates of each point from its image coordinates x_k, y_k and 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:

$$X_k = \frac{-Z_k}{f(x_k - x_0 + \delta_x)}, Y_k = \frac{-Z_k}{f(y_k - y_0 + \delta_y)}$$

where x_0 and y_0 are the coordinates of the principal point, and δ_x and δ_y are corrections for lens distortion [5]. The

[1]Kinect 2.0 documentation available on 6/4/2018 at: https://msdn.microsoft.com/en-us/library/jj131023.aspx

projection mask from the depth map space is transformed to the RGB image space using the obtained homography matrix. Geometric transformation of the depth map to the RGB image space using the inverse obtained homography matrix is visualized in Figure 4:

Figure 4: Geometric transformation of the depth map to the RGB image space using the inverse homography matrix of the RGB frame to the depth point-cloud space. The resolution of the frames provided by the Kinect device is different and, therefore, differs in the Figure, too.

The projection space of the LCD projector is relative to the RGB image space. Thus, we can compute another homography matrix between the RGB image space and the projection space. We use the projection plane corner coordinates to define a set of four equations with one unknown homogenous homography matrix H of size $3x3$:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = H \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

where projection plane corner coordinates defined earlier are substituted for x, 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 basic principles of algebra.

Output of the geometric perspective transformation applied on the projection mask in the RGB image space using the homogenous homography matrix is a projection mask in the projection space of the LCD projector. The resulting double transformed projection can be applied on the projectors frame buffer to project content directly on ROIs (Figure 3). Whole calibration process is repeated with all the LCD projectors.

4.2 Data acquisition

We used the SMI Eye Tracking Glasses[2] with 3-point calibration process[3] provided with SMI SDK for Windows platform[4] to acquire the egocentric video with corresponding gaze data during the user studies. The existing SMI SDK has two parts:

- server-side acquiring raw data from the eye tracker and sending the data to the client,
- client-side receiving and processing the data from the eye tracker.

We implemented an extension for the SMI SDK client-side. The extension is able to save the egocentric video from the eye tracker in a file with common, as loss-less as possible, video format (we used *wmv* video format). Moreover, the extension of the client-side saves the raw gaze data for each video frame in a separate structured *csv* file. Therefore, further pre-processing of the gaze data itself is possible in the next phase.

Pre-processing phase before the data evaluation is necessary. We adapted and implemented part of the I-VT fixation classifier algorithm published by Olsen [8]. We chose only suitable parts of the algorithm for our pre-processing procedure as the paper is only partially relevant for the mobile glass eye trackers:

- gap fill-in algorithm using linear interpolation,
- noise reduction algorithm using a median filter with small window size (3 gaze samples) as proposed by Olsen [8].

The pre-processed file with gaze data is an input for further data processing in the automatic evaluation phase.

4.3 Data evaluation

The data evaluation software module is proposed to make the evaluation of the information obtainable from the dataset simpler and faster. The evaluation output itself is designed to be universal, simple and useful for the future research with various objectives. For every frame of the egocentric video it consists of:

- the information about the content projected on ROIs at the moment,
- the index of the ROI on which an observer fixed his attention, if in any.

The evaluation output may be enhanced with some supplementary information, if needed in the future (i.e. delay of the first fixation after change of content on the ROIs). However, we do not take them into account for now.

The information we lack to produce the evaluation output are the coordinates of the ROIs in the egocentric video. We propose two methods of ROI segmentation from the egocentric video: segmentation of the image regions with the highest contrast and enhanced segmentation using fiducial markers [3] which was not evaluated, yet.

Segmentation of the image regions with the highest contrast is a simple and fast segmentation algorithm. The algorithm has an assumption that bright light (white colour)

[2]Manufacturer provides support for products no more. Available on 24/02/2018 at: `https://www.smivision.com/wp-content/uploads/2017/05/smi_prod_ETG_120Hz_asgm.pdf`

[3]Explanation of 3-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`

[4]Available on 24/02/2018 at: `https://www.smivision.com/wp-content/uploads/2016/10/smi_prod_sdk.pdf`

is projected on all ROIs at a frame in which the segmentation takes place. Therefore, the lightened ROIs are the brightest regions in the frame. The algorithm consists of two steps: image thresholding with high constant threshold value followed by the contour detection algorithm described by Suzuki [10] where detected contours represent the ROI border coordinates in the egocentric video.

Fiducial markers are used in computer vision to precisely detect coordinates of the marker position in an image. Therefore, the segmentation algorithm should be more reliable when fiducial markers projected on the ROIs are segmented from the egocentric video frame. This step is followed by the proposed segmentation of the image regions with the highest contrast to determine border coordinates of the ROIs. This algorithm was not used during our first user studies because this method is not indifferent on marker deformations which do not preserve straight lines. Therefore, marker content projected on the polystyrene balls will be undetectable.

We decided to keep track of the ROI locations between two consecutive frames using dense optical flow (algorithm by Farnebäck [2]) instead of performing segmentation frame-by-frame which can be time consuming. Dense optical flow is used because of the egocentric video specifics- head tilts, shakes, moves, and various depth ranges in the scene. We compute the approximal shift for every ROI coordinates between consecutive frames as ROI motion vector v which can be summed up with the border coordinates of the ROI from the previous frame to obtain the new ROI coordinates in the succeeding frame:

$$C(x,y)' = C(x,y) + \vec{v}; \quad \vec{v} = \overline{(V - (V \cap \neg C_{surr}))}$$

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, $\vec{v}$ is an approximal shift vector of the ROI, C_{surr} is the union of C with C surroundings in a frame in an absolute distance of δ pixels from C and V is a motion vector matrix between current and succeeding frame calculated by the dense flow algorithm [2].We take into account C_{surr} instead of C in the computation of the ROI motion vector due to distorted values of dense optical flow inside ROI with homogeneous content. The shift approximation error cumulation from the ROI tracking phase is eliminated every n-th video frame by repeated segmentation phase described in previous paragraph.

Fixated ROI at the certain frame is the one intersected by the gaze coordinates. Automatically evaluated frame by our evaluation module is captured in Figure 5. The area around the ROI throughout the evaluation process is enlarged due to the random error of the gaze coordinates from the eye tracker which was not documented, yet, but is significant in many cases.

4.4 Implementation details

We implemented proposed software modules in *C++* language with dependencies on several libraries:

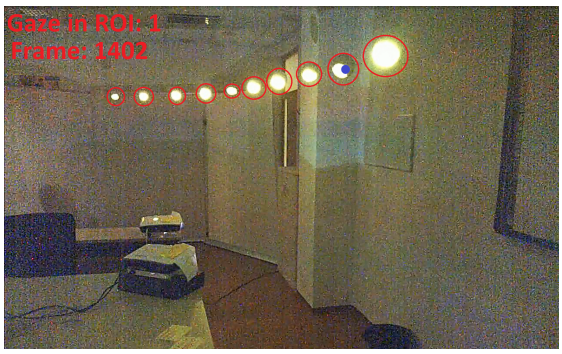

Figure 5: Automatically evaluated frame from the eye-tracker. Tracked ROIs are in red circles, gaze point in the current frame is a blue dot. Evaluation module stated that a fixation in ROI number 1 is present (numbered from 0, beginning from the right-most ROI).

- OpenCV library (version 3.4) including the OpenCV contribution package for the image processing and computer vision algorithms,
- GLFW 3 library for the operations on a graphic chip of the computer used in the projection handling module,
- Kinect for Windows Runtime v2.0 driver and Kinect for Windows SDK v2.0 for acquiring data from Kinect 2.0
- SMI SDK as the driver for the eye tracker

We obtained SMI SDK without any documentation from the digital appendix of Olesova's master thesis [7] as the SMI manufacturer provides no more support for its products since 2017.

5 Results

We created a novel dataset (contents described in the Section 3) from the data obtained by conducting the user studies with 37 participants based on our method proposal described in previous sections. Part of the information carried by the dataset was evaluated during studying the depth influence on the human visual attention. We evaluated observer's first fixation on certain ROI immediately after projection change on the scene to introduce the first results of our research. 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. However, these can be included in the conclusions of the future research. Moreover, we take into account only the first part of the projection sequence during the user studies described in Section 3 (projection of a white colour on two ROIs at the same time). Further parts of the projection sequence in the egocentric videos are in evaluation process at this time.

Statistical evaluation of the data has been made. Each ROI has been compared with every other in the means of the first fixation ratio after the concurrent change on the scene. Example of such a comparison of the ROI in the depth of 7,05 meters with every other ROI is visualized in Figure 6.

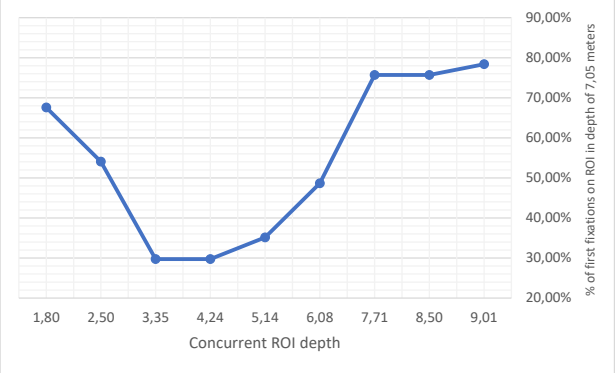

Figure 6: First fixations percentage after the concurrent changes on two ROIs on the scene at the same time occurred. In this example one ROI has depth of 7,05 meters and the others are situated in depths corresponding to the values on the x-axis.

We can compute the score of each ROI in comparison to every other from these first fixations percentages. Taking into account that concurrent changes on two ROIs are evaluated in the statistics, we claim that the first fixations percentage of 50% in the statistics (as in Figure 6) refers to pure coincidence (each of the two ROIs has attracted 50% 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 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. On the other hand, in the case of percentage 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.

We can express this relation as a saliency coefficient relative to the distance of the ROI from the observer. This can be done by normalization of first fixations percentage to the interval of $< 0; 1 >$ and computation of score for every normalized percentage value $percentage_{norm}$ using the equation:

$$score = percentage_{norm} - 0.5.$$

For each of the ROI we obtain the score from the interval $< -0.5; 0.5 >$ in comparison to every other ROI in different depths. Collision table of these coefficients for each ROI is visualized in Figure 7.

meters	9,01	8,50	7,71	7,05	6,08	5,14	4,24	3,35	2,50	1,80
9,01		-0,23	-0,392	-0,284	-0,122	-0,284	-0,311	-0,068	-0,257	-0,222
8,50	0,23		-0,149	-0,257	-0,203	-0,392	-0,311	-0,203	-0,311	-0,041
7,71	0,392	0,149		-0,257	-0,338	-0,257	-0,122	-0,176	-0,203	-0,014
7,05	0,284	0,257	0,257		-0,014	-0,149	-0,203	-0,203	0,041	0,176
6,08	0,122	0,203	0,338	0,014		-0,122	-0,365	0,23	0,203	0,041
5,14	0,284	0,392	0,257	0,149	0,122		-0,149	0,149	0,122	0,068
4,24	0,311	0,311	0,122	0,203	0,365	0,149		0,311	0,176	0,23
3,35	0,068	0,203	0,176	0,203	-0,23	-0,149	-0,311		0,176	0,149
2,50	0,257	0,311	0,203	-0,041	-0,203	-0,122	-0,176	-0,176		0,203
1,80	0,222	0,041	0,014	-0,176	-0,041	-0,068	-0,23	-0,149	-0,203	

Figure 7: Collision table of the depth saliency coefficients comparing ground-truth saliency of the objects, obtained during the user studies, in various depts with each other.

We can make some conclusions based on the analysis of the provided visualization. The most important conclusion is that depth plays significant role as an aspect of the human visual attention and further research of its impact on the visual attention is relevant. We can prove this claim by the fact that values of the coefficients vary and are distributed in the whole interval $< -0.5; 0.5 >$. Moreover, we can observe some trends in the coefficient values from the visualization. Coefficients under the slant are positive from the left-hand-side. This means that the objects closer to the observer have higher saliency in relation to their depth than the distant ones. In other words, the saliency of the objects on the scene was decreasing with their distance from the observer. However, this is not true about the objects too close to the observer. This leads us to an assumption, that human visual attention is perceiving certain depth of the scene with the highest priority (assigning it the highest saliency). We can call this the most salient depth. The saliency of the objects is then decreasing with the distance of the objects from the most salient depth of human visual attention. This assumption is supported by the Figure 8 containing the average coefficients of all the ROIs in different depths computed from the collision matrix (Figure 7).

6 Conclusions

We proposed a novel method to conduct experiments for user studies on visual attention in real world environments in a laboratory and introduced a new approach to study human visual attention on a real scene with possibility of dynamic changes in the scene. We held the first experiments following our proposed method in a laboratory with 37 participants to verify relevancy of the novel proposed method and to create a novel dataset for the research of human visual attention in real environments. We focused our research on studying the position of the objects on the scene (namely their distance from the observer) on their saliency.

We provided the first results by the statistical evaluation of the first fixations of the observers after dynamic changes on the scene. Results of the evaluation are visualized and discussed. Our research supports the claim that

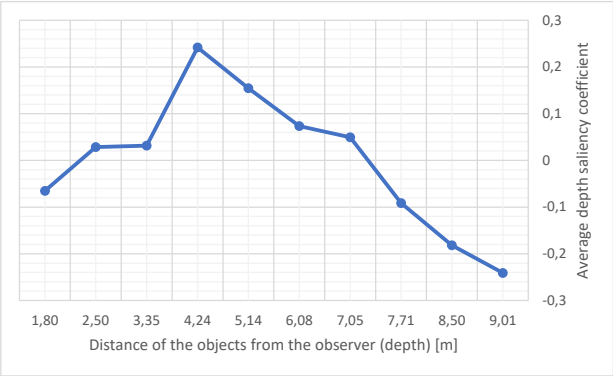

Figure 8: Average depth saliency coefficient of the ROIs in different depths computed from the ground-truth obtained during the user studies. The most salient depth is approximately in the distance of 4-5 meters from the observer. The saliency of the objects decreases significantly with the distance from the most salient depth.

depth plays significant role as an aspect affecting human visual attention and can be approximated as a saliency coefficient applicable on an existing model of human visual attention. Moreover, we claim that there exists the most salient depth: the range of the distances from the observer where objects have the highest saliency for the observer. The saliency then decreases significantly with the distance from the most salient depth. We assume that applying our knowledge on an existing model of visual attention will provide accurate model of human visual attention and will better reflect saliency of the objects in a real-world environments. However, further research using our proposed method is necessary to prove our assumption.

The novel dataset is still in evaluation process at this time. Future work related to this paper will cover evaluating results from the extensive user studies and creating an output robust enough to approximate the depth influence on human visual attention as a weighting function applicable on an existing saliency models. We will evaluate the models enhanced with our knowledge and will publish our results.

Acknowledgement: The author would like to thank for financial contribution from Slovakian Grant VEGA 1/0874/17, UVP Project STU and from the Research and Development Operational Programme for the project University Science Park of STU Bratislava, ITMS 26240220084, co-funded by the European Regional Development Fund, and e-Talent Nadacie Tatrabanky 1918.

References

[1] Ali Borji, Laurent Itti, J Liu, P Musialski, and P Wonka. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[2] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.

[3] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[4] E Bruce Goldstein. *The Blackwell handbook of sensation and perception*. John Wiley & Sons, 2008.

[5] Kourosh Khoshelham and Sander Oude Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012.

[6] Congyan Lang, Tam V Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan. Depth matters: Influence of depth cues on visual saliency. In *Computer vision–ECCV 2012*, pages 101–115. Springer, 2012.

[7] Veronika Olesova. Generating a Saliency Map using Depth Information. Master's thesis, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Slovakia, 2016.

[8] Anneli Olsen. The tobii i-vt fixation filter. *Tobii Technology*, 2012.

[9] Katherine L Roberts, Harriet A Allen, Kevin Dent, and Glyn W Humphreys. Visual search in depth: The neural correlates of segmenting a display into relevant and irrelevant three-dimensional regions. *NeuroImage*, 122:298–305, 2015.

[10] Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 30(1):32–46, 1985.

[11] Junle Wang, Matthieu Perreira Da Silva, Patrick Le Callet, and Vincent Ricordel. Computational model of stereoscopic 3d visual saliency. *IEEE Transactions on Image Processing*, 22(6):2151–2165, 2013.

[12] Mai Xu, Yun Ren, and Zulin Wang. Learning to predict saliency on face images. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 3907–3915. IEEE, 2015.

[13] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 28–31. IEEE, 2004.