Veronika Olesova, Wanda Benesova and Patrik Polatsek, " Visual Attention in Egocentric Field-of-view using RGB-D Data," Proc. SPIE 10341, Ninth International Conference on Machine Vision (ICMV 2016), 103410T (March 17, 2017)

http://dx.doi.org/ doi:10.1117/12.2268617

Visual Attention in Egocentric Field-of-view using RGB-D Data.

Veronika Olesova, Wanda Benesova, Patrik Polatsek
Slovak university of technology, Faculty of Informatics and Information Technologies

ABSTRACT

Most of the existing solutions predicting visual attention focus solely on referenced 2D images and disregard any depth information. This aspect has always represented a weak point since the depth is an inseparable part of the biological vision. This paper presents a novel method of saliency map generation based on results of our experiments with egocentric visual attention and investigation of its correlation with perceived depth. We propose a model to predict the attention using superpixel representation with an assumption that contrast objects are usually salient and have a sparser spatial distribution of superpixels than their background. To incorporate depth information into this model, we propose three different depth techniques. The evaluation is done on our new RGB-D dataset created by SMI eye-tracker glasses and KinectV2 device.

Keywords: saliency map, visual attention, egocentric video, RGB-D data, eye-tracker glasses

1. INTRODUCTION

Visual attention has been studied in many research areas including machine learning, computer vision, psychology or signal processing. Understanding where people look in the scene is very useful in applications of image processing such as image compression, scene interpretation, and computer graphics but also in marketing. We can predict the visual attention by various models, that typically represent this information in the so-called visual saliency maps. Most existing saliency models investigate only cues from 2D images (color, orientation, luminance or texture), which might lead to inaccurate saliency detection since our visual system operates in 3D environments. Two main factors contribute to the decision whether the subset in the visual field is salient or not. The first one, bottom-up saliency, depends only on the instantaneous sensory input, without taking into account the internal state of the organism [1]. It is fast and involuntary. In contrast to this, top-down saliency that takes into account the internal state, is slow, task-driven and voluntary. The methods described below deal mainly with the low level of attention and bottom-up saliency models.

2. VISUAL ATTENTION AND SALIENCY MODELS USING RGB-D DATA

There have been proposed various approaches to compute saliency, such as hierarchical, Bayesian, decision-theoretical, information-theoretical, graphical, using spectral analysis or pattern classification. One of the most known bottom-up hierarchical model is proposed by Itti et al. [1]. It extracts three visual features: color, intensity and orientation. Normalised feature maps are combined into three conspicuous maps for intensity, color and orientation and finally into a single saliency map.

Our approach is focused mainly on the hierarchical superpixel-based models. Superpixels are regions in an image which can be used as basic units (primitives) in the next image processing like segmentation, salience mapping or object detection [2].

2.1 Visual attention and 3D visual features

Following studies have examined how visual attention may be influenced by 3D visual features and analyzed the difference between 2D and 3D eye fixation data. Lang et al. in their work [3] investigated whether the spatial distributions of eye fixations differ for 2D and 3D images. The authors collected a larger eye fixation dataset for 2D-vs-3D scenes. Each participant was assigned two blocks of 100 randomly chosen images, one of the blocks contained 2D while the other 3D images. Both blocks were viewed on a 3D LCD display in the corresponding mode (3D or 2D) using the active shutter glasses in case of 3D images. The eye fixations were captured using an infra-red illumination based remote eye-tracker. The images were displayed in random order for 6 seconds followed by a grey mask for 3 seconds. The observations coming from their study are as follows:

- Depth cues modulate visual saliency to a greater extent at farther depth ranges. Furthermore, humans fixate preferentially at closer depth ranges.
- A few interesting objects account for majority of the fixations and this behavior is consistent across both 2D and 3D.
- The relationship between depth and saliency is non-linear and characteristic for low and high depth-of-field scenes.
- The additional depth information led to an increased difference of fixation distribution between 2D and 3D version, especially, in case of multiple salient stimuli located in different depth planes.

Ma CY and Hang HM published in [4] their learning-based saliency model with depth information. They have also collected a 3D eye fixation dataset, with a slightly different experimental setup. They used Tobii TX300 eye- tracker and a 23-in patterned retarder 3D display. The experiments did not show a significant difference between watching 2D and 3D content, except for the first three fixations. In these initial fixations people look at the most interesting objects. However, in the next fixations their eyes move to other areas in the image and the 2D low-level features dominate human visual attention again.

Jansen et al. [5] investigated the influence of disparity on viewing behavior in the observation of 2D and 3D still images. They found that the additional depth information led to an increased number of fixations, shorter and faster saccades, and broader spatial exploration. However, no significant difference was found between the viewing of 2D and 3D stimuli concerning the saliency of several 2D visual features (mean luminance, luminance contrast and texture contrast).

Hakkinen et al. [6] measured and compared the eye movements of participants watching the same video in 2D and 3D versions. The result of this research shows, that eye movements are more widely distributed in 3D content. Viewers watching 3D content paid attention also to other targets than main actors (those were the only target in 2D content). The main contribution of this work is the fact, that depth provides additional information about scene and thus creates new salient areas. This result suggests the existence of a saliency map from depth, and a potential summation operation during the integration of 2D and depth saliency information.

Wang et al. [7] found that objects closest to the observer always attract the most fixations. The number of fixations on each object decreases as the depth order of the object increases, except for the furthest object which receives a few more fixations than the one or two objects in front of it. Considering the influence of center-bias in 2D visual attention, these results indicate the existence of an additional location prior according to the depth in the viewing of 3D content. This location prior indicates the possibility of integrating depth information by means of a weighting.

2.2 Methods of an incorporation of depth information into the model

Hence, the depth has an impact on the visual attention and incorporating the depth information into the saliency model is reasonable. Existing computational models that use depth information can be classified into three categories [8]:

Depth-Weighting Models [9], [10], [11]: These models treat a depth information as a weighting factor and does not contain any feature-extraction process. The saliency value of each pixel of the resulting map is directly related to its depth. In addition to 2D scene we also need depth map as the input. This map can be acquired either by depth detection equipment (e.g. Kinect device) or using depth estimation algorithm on two views.

The model of Cheng et al. [15] modifies an existing approach based on region contrast by adding a depth feature. The depth information is in this model treated only as a weighting factor.

Depth-Saliency Models [12], [13]: This category of models takes depth saliency as additional information. They first extract depth features from the depth map to create feature maps. The depth saliency maps are then generated and combined with 2D saliency maps.

Wu et al. [16] proposed a model of RGB-D salient object detection via feature fusion and multi-scale enhancement. They first convert the input RGB image into CIE L*a*b space and normalize the depth image into the range [0, 255]. The converted color image is then segmented into superpixels. Two adjacent superpixels are merged together if the difference between their average depths is under a threshold value. Computed color and depth contrast for each merged superpixel is subsequently fused to create a saliency map. Finally, the multi-scale enhancement is applied to this map to improve the detection precision.

Stereo-Vision Models [14]: The last type of models takes into account the mechanisms of the stereoscopic perception in the human visual system. Unlike the first two models, this model takes two stereo images as input, from which 2D visual features can be considered.

3. OUR CONTRIBUTION

3.1 Experimental studies on visual depth attention

The goal of the first experiment was to find out: How does distance between observer and observed object (depth) affect the visual attention? The experimental scene was constructed in the school laboratory, where we have evenly distributed and hung twelve objects wrapped in colored paper on the wall. Red, beige and yellow colors were used for the wrapping paper. The farthest object was located approximately 9 meters from the participant. The setup is illustrated in Figure 1.

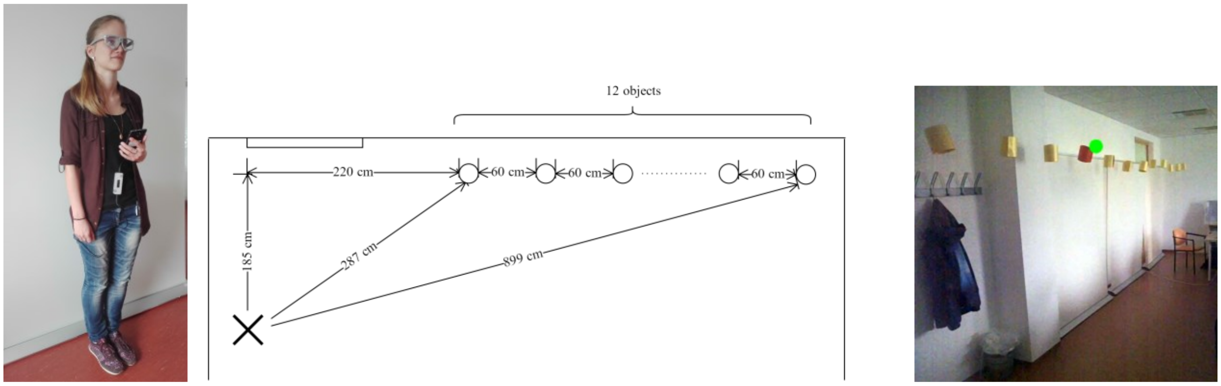


Figure 1. CroParticipant with SMI eye-tracker glasses (left), experimental scene aquired by SMI eye-tracker glasses (middle) and layout of experimental scene where cross represents position of the participant while objects are illustrated as circles (right).

In the described experimental scene, we have done three experiments using eye-tracking glasses.

- **Experiment 1**: all 12 objects (see Figure 1) were beige (L*a*b* = 80, 8, 50),
- **Experiment 2**: one object was red (L*a*b* = 38, 61, 39) and the rest were beige (L*a*b* = 80, 8, 50). The depth position of the red object has been changed during the experiment, always for the next participant.
- **Experiment 3:** one object in variable position was yellow (L*a*b* = 81, 13, 97) and the rest were beige (L*a*b* = 80, 8, 50). In this case, the depth position of the yellow object has been changed during the experiment, always for the next participant. (All L*a*b* values are for daylight D65.)

We had 20 participants, each of them has participated in all three experiments after a 3-point calibration. Consequently, they were instructed to walk into the room and stand on the cross position. Participants had to freely look at the scene without any specific task given. From each observation of scene, we took only the first three fixations into account, since for some participants there were no more fixations or after this number the fixations were repeated. In the process of evaluation, we have assigned a greater weighting to the first fixation than the third one to account for the order of these fixations. To evaluate the first experiment, we have summed up every weighted fixation from all the participants at each object separately, resulting in Figure 2. From the graph we can see that people tend to look at the closer objects sooner than those in the back. Evaluation of the second experiment indicates that as long as the red object was at one of the first 9 positions, it drew attention of the most participants. Such object at the farther locations does not seem to have an impact on the human attention. Evaluation of the third experiment brought a different observation. Despite the small number of participants, the measured gaze positions highlight the fact that this color does not have such an evident impact on human attention as the red object in the second experiment. The highlighted object does not grab the participants' attention. Their eye fixations have a random distribution, mainly around the objects closer to the participant. The color difference between beige and yellow objects was L*a*b* = 46 what is considerably lower than the color difference between beige and red objects in the second experiment, which was: L*a*b* = 68. In Figure 2 we can see also the result of fitting process, where the actual distance from object is used and the farthest object is removed because it exceeded Kinect range. After the normalization we got the following function:

$$y = 2*10^{-7}x^3 - 0.0002x^2 + 0.0287x - 0.7544 \quad (1)$$

A 2D saliency map is then refined by multiplying each value by output of this polynomial function.

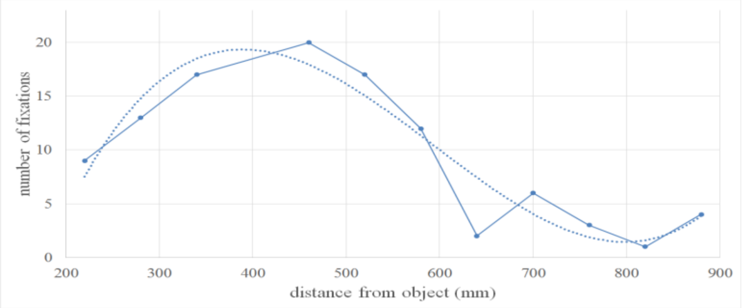


Figure 2. Gaussian scale-space pyramid create an interval in the difference-of-Gaussian pyramid.

3.2 New video dataset consisting of RGB-D Data and gaze position from eye-tracker glasses

The second set of experiments was done in a complex scene using the same eye-tracker with an additional Kinect device mounted on the head. The setup is illustrated on the Figure 3. We have placed many different sized and colored objects like flowers, books, toys etc. in the experimental scene, some closer than the others. Unlike the previous experiments, the participant could freely move around the room and look at any object that attracts his attention. Our final dataset includes experimental data of 6 participants, having video records in the range of 15 to 30 seconds. While Kinect device has a frame rate of 30 fps, the SMI eye-tracker has only 24 fps. To properly synchronize their outputs, we store also timestamp of each frame.

Saliency model expects 3 images on the input: color, depth and fixation image. So far we have color and depth images from the Kinect device, but not the fixation image. The transformation of gaze information from eye-tracker coordinate system to Kinect is therefore necessary. Corresponding pairs of color frames for the whole video is found automatically using timestamps, we just need to manually find a single initial pair. To get a location of a gaze point in Kinect space, we create a homography between these pairs.

3.2.1 Image registration

Each of the pair of frames in videos acquired simultaneously by Kinect and also by the eye-tracker camera should be registered. The homography can be found by providing the corresponding key-points of two images. We have achieved the best results using SIFT algorithm [17] to detect and describe local features in a combination with Brute-Force matcher (cross check enabled). Looking at pairs of images we noticed that an image from the eye-tracker will be always a subset of a Kinect image and the whole scene will have some small offset due to different angles of devices. Based on this assumption, we kept only matches whose distance of x coordinates (y coordinates) did not exceed the mean (median) of all the other matches. The homography is then found using the RANSAC method [18] (Figure 3).

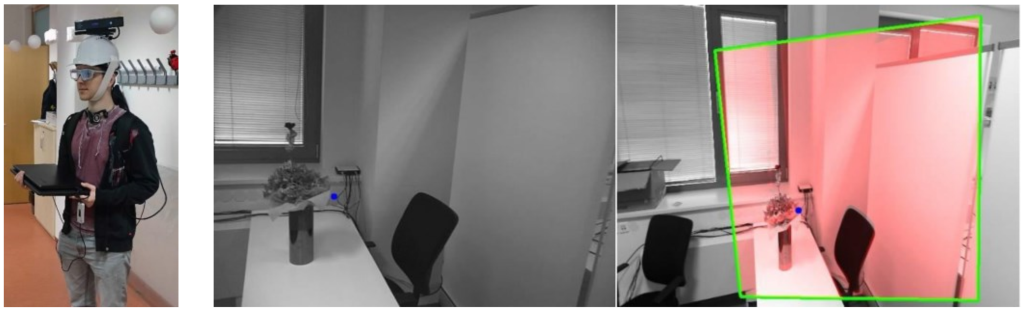


Figure 3. Experimental setup of the second series of experiments (left), Visualization of a homography: eye-tracker image (middle), Kinect image (right). Blue dot is the measured gaze position.

3.2.2 Post-processing of the depth images

Zero values in the depth data indicate that the objects depth could not be estimated. In addition to missing depth values caused by object being out of range, in our depth frame we had another artefact in the form of black borders on left and right side of the image (Figure 4a). This is a consequence of our previous mapping of depth frame to color space that has to be done because of different angles of sensors. For the further use of these data we had to supplement the missing values. First of all, we have removed the salt and pepper noise using the median blur as shown in Figure 4b. Then, we have created a binary mask (Figure 4c) representing the invalid pixels. To restore the values of the mask region we used an in-paint method that calculates the missing values restricted by the binary mask using the neighborhood. The final smoothened depth can be seen in Figure 4d.

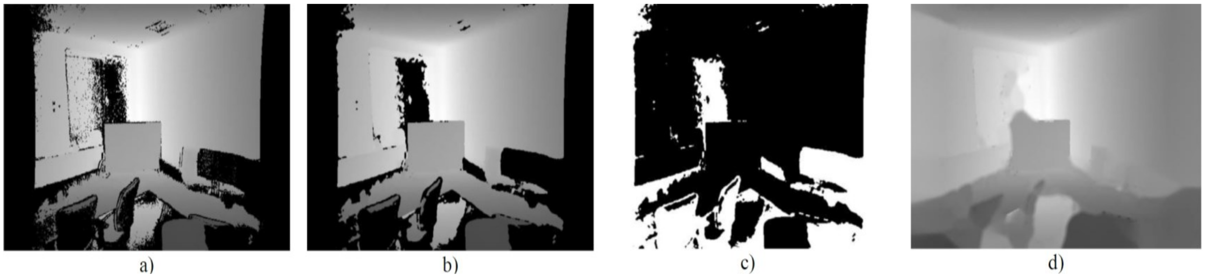


Figure 4. Depth smoothing: a) raw depth image, b) removed salt and pepper noise, c) noise mask, d) smoothened depth image.

3.2.3 Egocentric RGB-D eye-tracker dataset

The final version of the created dataset RGB-D Gaze consists of 6 egocentric videos acquired by individual participants. Each of them includes RGB-frame, D-frame and the corresponding measured gaze position in each frame. Number of frames in one video varies in the range of 570 -1100. Dataset is available for download on the web site:
http://vgg.fiit.stuba.sk/2016-06/egocentric-rgb-d-eye-tracker-dataset/

4. PROPOSED SUPERPIXEL-BASED SALIENCY MODEL USING RGB-D DATA

Our aim is an enhancement of a 2D saliency model by the depth information using RGB-D data. In general, the proposed incorporation of the depth information should be also adapted for any 2D saliency model. For this evaluation, we have implemented the 2D superpixel-based saliency model as described in [19], which computes global contrast and spatial spread for each superpixel. The assumption is that salient object superpixels usually show noticeable color contrast with background superpixels and the spatial distribution of salient object superpixels is sparser than background superpixels.
To incorporate the depth information into the saliency model, we have proposed three different techniques that are described in this section.

Depth contrast (DC): In the first step, a calculation of superpixels in the depth image has been done. Depth contrast is computed as a difference between mean depths of two superpixels:

$$DC(i) = \sum W(i,j) \cdot \left\| md_i - md_j \right\| \qquad (2)$$

where md_i stands for the mean depth of the i-th superpixel and the weight W(i, j) is defined as:

$$W(i,j) = \left| SP_j \right| \cdot Sim_d(i,j) \qquad (3)$$

where $|SP|_j$ stands for the number of pixels in the superpixel and $Sim_d(i, j)$ is the distance between two superpixels. This feature DC(i) is added to the final multiplication of the color contrast and spatial spread to generate the final RGB-D saliency map.

Simple (linear) depth Weighting (SD): We have simply divided each saliency value by the pixel's depth. At this time, we supposed that human attention is not just linear and this approach would not lead to good results, which is what we wanted to prove.

Advanced Depth Weighting (AD): Our intention in the last technique was to take into account the actual human depth perception. We made use of our first experiment with eye-tracker glasses and the fitted polynomial curve (Figure 2). In this case, the polynomial function was used as a multiplicative factor for the depth information.

4.1 Implementation of the experimental system

The experimental system has been implemented in the C++. A diagram of the whole calculation process is presented in Figure 5.

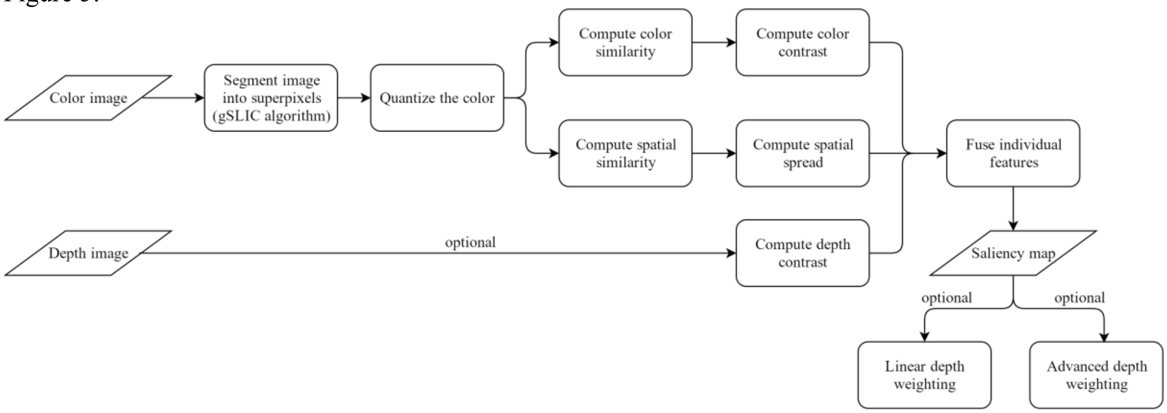


Figure 5. Diagram of the implemented experimental system.

5. EVALUATION AND RESULTS

5.1 Examples of the evaluation images

In the following examples (Figure 6) we compare results of saliency map with depth contrast (DC) and no depth cases (ND) with real fixations. We have labelled image locations that received the highest saliency value with blue (for depth contrast) and red circles (for no depth). A human fixation is represented by a green circle in the image.

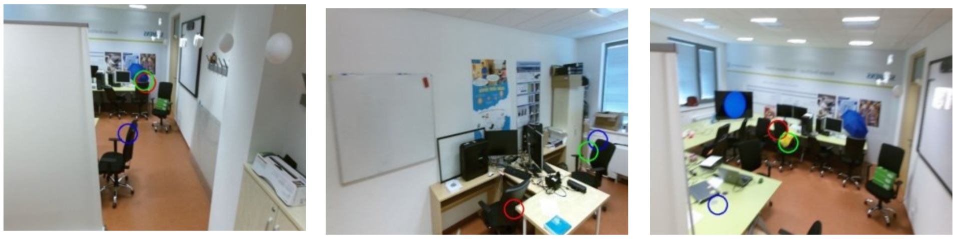
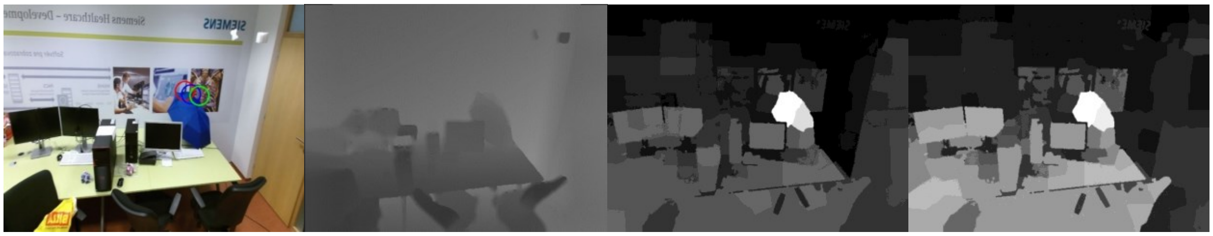


Figure 6. Examples of results with different impacts of the depth information.

In Figure 7 we provide a comparison of different saliency maps of a single frame for a better understanding of visual differences between depth incorporation techniques that we have implemented.



a) b) c) d)

Figure 7. Contrast object in the scene: a) original image with labels, b) depth image, c) saliency map with no depth, d) saliency map with depth contrast.

5.2 Results

We have evaluated the following 4 cases of the depth incorporation into the previous saliency model:

- no depth (ND),
- depth contrast (DC),
- simple (linear) depth weighting (SD),
- advanced depth weighting (AD).

ROC (Receiver Operating Characteristic) curves (Figure 8 left) of the proposed depth techniques shows that the linear weighting produces the worst results. Differences between other depth cases are very small and difficult to distinguish from ROC curves and AUC (Area Under Curve) metric. However, the NSS (Normalized Scanpath Saliency) comparison in (Figure 8 right) revealed that depth contrast achieves the best results and together with advanced depth weighting performs better than case with no depth included.

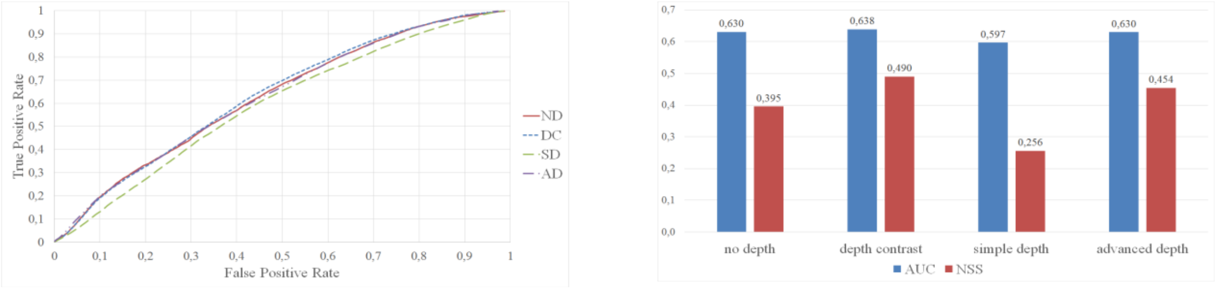


Figure 8. ROC curves of different cases of depth incorporation (left) and AUC and NSS comparison of depth incorporation (right).

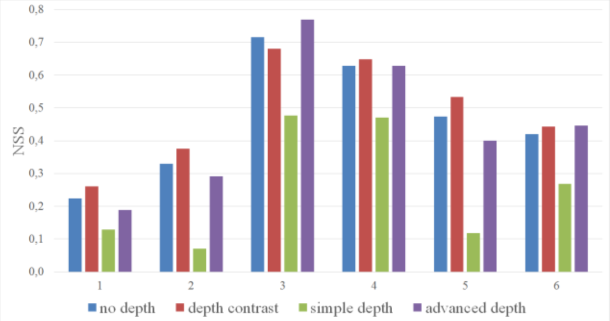


Fig. 9. Comparison of the depth impact on each individual participant.

To compare results between participants, we have created a clustered column chart (Figure 9), where each cluster represents one individual. The chart underlies the theory that each individual perceives color, size and depth of objects in a slightly different way. While attention of participant with number 1 and 2 does not match with our saliency model very well, we were able to predict attention of participant 3 and 4 reasonably. There was no participant who would perceive depth in a linear manner (linear depth). On the other hand, depth contrast could predict attention almost of every participant most reliable. Our hypothesis is that the visual attention model should be adaptive for various type of observers. In the future work, we will strongly focus on this problem.

6. CONCLUSION

Lack of published 3D datasets is the main problem of modeling 3D visual attention. We have therefore created an innovative dataset captured from a first person-view perspective containing RGB-D images. The setup that we used consists of SMI eye-tracker glasses and the Kinect device. We have conducted several experiments with human visual attention and came to the conclusion that depth has an influence on our attention. Participants were generally looking at larger and more contrast objects, another very important feature was the depth itself. We also observed that attention strongly depends on an individual participant although all of them were instructed in the same way. Our saliency model uses superpixels as the basic units of image space and computes their global color contrast and spatial spread in order to predict human attention. To improve its results, we have implemented three techniques of depth incorporation. First one, depth contrast, promoted superpixels with a unique depth compared to the rest of superpixels. Linear depth technique promoted pixels linearly assigned higher values to objects closer to camera, lower to further ones. This linear depth technique was not sufficient to predict human attention. The last one, advanced depth technique, was based on additional depth contrast feature and advanced depth technique led to better results of the saliency model. This model was evaluated on our new created dataset.

In the future work, we plan to extend the existing experiment in order to evaluate more participants. Moreover, we want to prove that there is a possibility of the data clustering and, finally, we want to create an adaptive attention model using the RGB-D data.

ACKNOWLEDGMENT

The authors would like to thank for financial contribution from the STU Grant scheme for Support of Young Researchers, to Slovakian Grant VEGA 1/0625/14 and to 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.

REFERENCES

[1] Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis & Machine Intelligence (11) (1998) 1254-1259

[2] Polatsek, P., Benesova, W.: Bottom-up saliency model generation using superpixels. In: Proceedings of the 31st Spring Conference on Computer Graphics, ACM (2015) 121-129

[3] Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: Inuence of depth cues on visual saliency. In: Computer Vision-ECCV 2012. Springer (2012) 101-115

[4] Ma, C.Y., Hang, H.M.: Learning-based saliency model with depth information. Journal of vision 15(6) (2015) 19-19

[5] Jansen, L., Onat, S., König, P.: Inuence of disparity on _xation and saccades in free viewing of natural scenes. Journal of Vision 9(1) (2009) 29-29

[6] Häkkinen et al.: What do people look at when they watch stereoscopic movies? In: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics (2010) 75240E-75240E

[7] Wang, J., Le Callet, P., Tourancheau, S., Ricordel, V., Da Silva, M.P.: Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli. Journal of Eye Movement Research 5(5) (2012)

[8] Wang, J., DaSilva, M.P., LeCallet, P., Ricordel, V.: Computational model of stereoscopic 3d visual saliency. Image Processing, IEEE Transactions on 22(6) (2013) 2151-2165

[9] Maki, A., Nordlund, P., Eklundh, J.O.: A computational model of depth-based attention. In: Pattern Recognition, 1996., Proceedings of the 13th International Conference on. Volume 4., IEEE (1996) 734-739

[10] Zhang, Y., Jiang, G., Yu, M., Chen, K.: Stereoscopic visual attention model for 3d video. In: Advances in Multimedia Modeling. Springer (2010) 314-324

[11] Chamaret, C., Gode_roy, S., Lopez, P., Le Meur, O.: Adaptive 3d rendering based on region-of-interest. In: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics (2010) 75240V-75240V

[12] Ouerhani, N., Hügli, H.: Computing visual attention from scene depth. In: Pattern Recognition, 2000. Proceedings. 15th International Conference on. Volume 1., IEEE (2000) 375-378

[13] Potapova, E., Zillich, M., Vincze, M.: Learning what matters: combining probabilistic models of 2d and 3d saliency cues. In: Computer Vision Systems. Springer (2011) 132-142

[14] Bruce, N.D., Tsotsos, J.K.: An attentional framework for stereo vision. In: Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on, IEEE (2005) 88-95

[15] Cheng, M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.: Global contrast based salient region detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on 37(3) (2015) 569-582

[16] Wu, P., Duan, L., Kong, L.: Rgb-d salient object detection via feature fusion and multi-scale enhancement. In: Computer Vision. Springer (2015) 359-368

[17] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2) (2004) 91-110

[18] Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model _tting with applications to image analysis and automated cartography. Communications of the ACM 24(6) (1981) 381-395

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