International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

A region covariances-based visual attention model for RGB-D images

Erkut Erdem[1]*

Accepted 15th October 2016

Abstract: Existing computational models of visual attention generally employ simple image features such as color, intensity or orientation to generate a saliency map which highlights the image parts that attract human attention. Interestingly, most of these models do not process any depth information and operate only on standard two-dimensional RGB images. On the other hand, depth processing through stereo vision is a key characteristics of the human visual system. In line with this observation, in this study, we propose to extend two state-of-the-art static saliency models that depend on region covariances to process additional depth information available in RGB-D images. We evaluate our proposed models on NUS-3D benchmark dataset by taking into account different evaluation metrics. Our results reveal that using the additional depth information improves the saliency prediction in a statistically significant manner, giving more accurate saliency maps.

Keywords: *Visual attention, Visual saliency, Depth saliency, RGB-D images, Region covariances.*

1. Introduction

Amount of visual information captured through the eyes is so vast that the brain develops certain mechanisms to process them. Visual attention, as an umbrella term, denotes these mechanisms which are responsible from selecting the most relevant parts of the visual data while discarding the rest. This selection procedure is carried out by fixating the eyes to certain locations of the visual data. Visual attention is an umbrella term since it has been shown that it includes both bottom-up and top-down mechanisms. Bottom-up attention is mostly involuntary and processes in a purely data-driven manner. In short, it is in charge of extracting the image parts which have very different characteristics than their surroundings in terms of different visual features. On the other hand, top-down attention is task-specific and processes the data in a goal-dependent manner, and thus semantic and top-down knowledge play a key role in top-down attention.

In computer vision literature, researchers have developed several computational approaches for visual attention to predict where human look at images [4,5,8,10-12,14,18,20,24,28,29,32] (see Figure 1). Most of these visual attention models follow the same bottom-up architecture. By processing the raw visual data, they first extract certain visual features such as intensity, color and orientation, and then compute individual saliency maps for each one of these features. As the last step, they combine these individual maps (after applying certain normalization strategies) to output a final saliency map whose maxima indicate the image locations that attract human attention. For a detailed review of these models, the reader can refer to [2]. It is important to note that the recent trend in saliency prediction is to use deep learning, but to perform an end-to-end learning they need very large image datasets with eye fixation data (e.g., [21,26]).

Although, in recent years, we have witnessed a huge increase in the number of visual saliency models, most of these models operate on RGB images and do not use any depth information. On the other hand, the human visual system has a stereo vision capability which enables to capture and process the depth information. In that respect, it can be argued that these attention models do not fully mimic the human visual system. Motivated with this observation and the recent advances in the 3D-capable acquisition equipments such as RealSense and Microsoft Kinect, some researchers have developed a number of depth-aware visual saliency models [3,16,17,19,22,23,25,27,31].

In this paper, we propose two new visual saliency models that additionally process depth information and operate on RGB-D images. A RGB-D image is an image which consists of four channels with the first three channels forming a standard RGB image, and the last channel denoting a depth channel aligned with the RGB component. In particular, we extend the previously suggested CovSal image saliency models [8] to operate on RGB-D images. These models originally operate on RGB images and predict saliency by estimating the center-surround differences based on first and second order feature statistics. In our work, we employ these models to estimate two different saliency maps, one from the RGB image and the second from the depth image. We then combine these RGB and depth saliency maps and output a single saliency map for a RGB-D image. As the second contribution of the paper, we evaluate the effect of using depth information through these models in saliency prediction on the NUS-3D benchmark dataset [22] by using several different evaluation metrics. Our results demonstrate that using depth information in saliency estimation improves the prediction accuracy.

The paper organization is as follows: In Section 2, we provide a brief overview of the existing depth-aware saliency models. In Section 3, we review the CovSal models and present the proposed extensions to these models to process depth information through RGB-D images. In Section 4, we evaluate the proposed models and present our qualitative and quantitative results. Finally, in Section 5, we provide a summary and discuss our findings.

[1] *Hacettepe University. Department of Computer Engineering, Ankara, Turkey – TR-06800*
* *Corresponding Author: Email: erkut@cs.hacettepe.edu.tr*

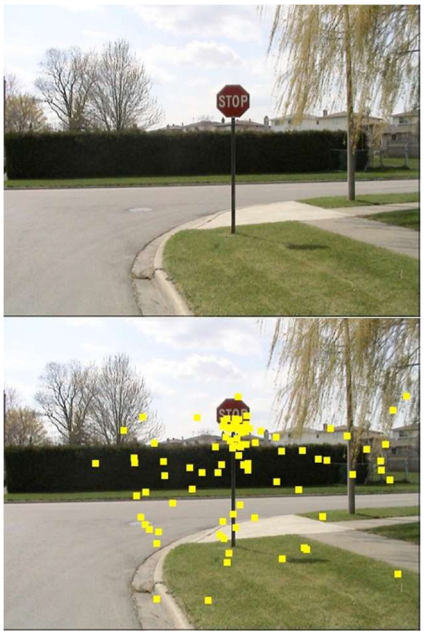

Figure 1. A sample image and the corresponding eye fixation data collected from the human subjects, representing where they look at the image (taken from [4]). The raw image is given at the top, and the same image with the superimposed fixations is given at the bottom.

2. Related Work

As mentioned in the introduction, most of the computational models of visual attention focus on 2D images which are generally represented as RGB images. Although these models have attracted much attention lately, predicting depth-aware saliency has received relatively little attention. There is a limited number of saliency models which process the depth information in a 3D environment [3,16,17,19,22,23,25,27,31]. These existing depth-aware models can be categorized into three groups in terms of how they employ the depth information in saliency prediction [15,31].

The first group of models, which are referred to as stereovision models, work directly on stereo images (i.e. left and right views of a scene), and use the binocular cues extracted from them as the additional features in saliency prediction without explicitly estimating the depth information [3,17]. For instance, Bruce and Tsotsos proposed such a saliency model in [3] by extending their 2D saliency model to include these kind of stereo image features. Iatsun et al. followed a similar approach in [17] and suggested to use binocular depth features in saliency estimation.

Compared to the first group, the second line of models, which are known as depth-weighting models, assume that a depth map is explicitly computed and available, and use these depth maps as a weighting factor for the saliency prediction [22,33]. Specifically, motivated by the observation that the observers in general attract to the regions that are closer to themselves and not the regions far away, these models compute a 2D saliency map from the input RGB image and then perform a point-wise multiplication of this saliency map and the given or the estimated depth map. An example model within this group is the saliency model proposed by Lang et al. in [22]. They learn a depth prior from a set of training images and employ it as the weighting factor for a 2D saliency model.

The last group of models, which are called depth-saliency models, either employ the explicitly provided depth image as an additional feature channel within a saliency model [15,16,19,23, 25], or extract a saliency map from the depth image alone which is then combined with the saliency map predicted from the RGB image [17, 31].

Our proposed approach is an example of the latter group of depth-saliency models. For a given RGB-D image, we extract both a traditional saliency map from the RGB channels and a depth saliency map from the aligned depth image. We then linearly combine these maps to generate the final saliency map. Our models employ the CovSal image saliency models [8] while extracting both of these saliency models, and thus they can be regarded as extensions of these previously proposed models which additionally consider the available depth information.

3. Proposed Approach

Our depth-aware saliency models depend on the CovSal saliency models [8]. Hence, in this section, we first review the computational details of the CovSal saliency models. Next, we discuss the proposed extension of these models to process extra depth information.

3.1. CovSal Saliency Model.

CovSal saliency models proposed by Erdem and Erdem in [8] are based on region covariances [30]. Motivated by the observation that region covariances encode local geometry (structure or texture) of an image region, the models define the center-surround differences, which are key for saliency estimation, in terms of the second order feature statistics given by the region covariance representation.

In particular, let I denote a 2D color image given as the input, and

$$F(x,y) = \left[L(x,y) \quad a(x,y) \quad b(x,y) \quad \left|\frac{\partial I(x,y)}{\partial x}\right| \quad \left|\frac{\partial I(x,y)}{\partial y}\right| \quad x \quad y\right]^T \tag{1}$$

represent the seven dimensional feature vector extracted at the image pixel at the pixel location *(x, y)*. Here, *(L, a, b)* denotes the color information represented in CIE La*b* color space, and $\partial I(x,y)/\partial x$ and $\partial I(x,y)/\partial x$ respectively represent the first order derivates of the intensity in horizontal and vertical dimension. Then, an image region of n by n pixels defined over the image I can be represented by the corresponding feature statistics $(\boldsymbol{\mu}, \boldsymbol{C}_R)$ where $\boldsymbol{\mu}$ denotes the mean feature vector and

$$\boldsymbol{C}_R = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{f}_i - \boldsymbol{\mu})(\boldsymbol{f}_i - \boldsymbol{\mu})^T \tag{2}$$

is the corresponding covariance matrix with $\{\boldsymbol{f}_i\}_{\,i=1\ldots7}$ being the 7 dimensional feature vector given in Equation (1).

The CovSal saliency models employ these first and second order feature statistics to estimate the saliency value of a given image region R_i centered at the pixel location x_i by using the following equation:

$$S(R_i) = (1 - \|x_i - x_c\|/Z)\,\Sigma_{i=1\ldots m}\left((1/m)\; d(R_i, R_j)\right) \tag{3}$$

In Equation (3), the first term $(1 - \|x_i - x_c\|/Z)$ is the weighting factor introduced for the center bias and lets the regions close to the image center x_c have higher saliency values than the others. Here, $Z = max_{i^* \in I}\|x_{i^*} - x_c\|$ is used as a normalization factor.

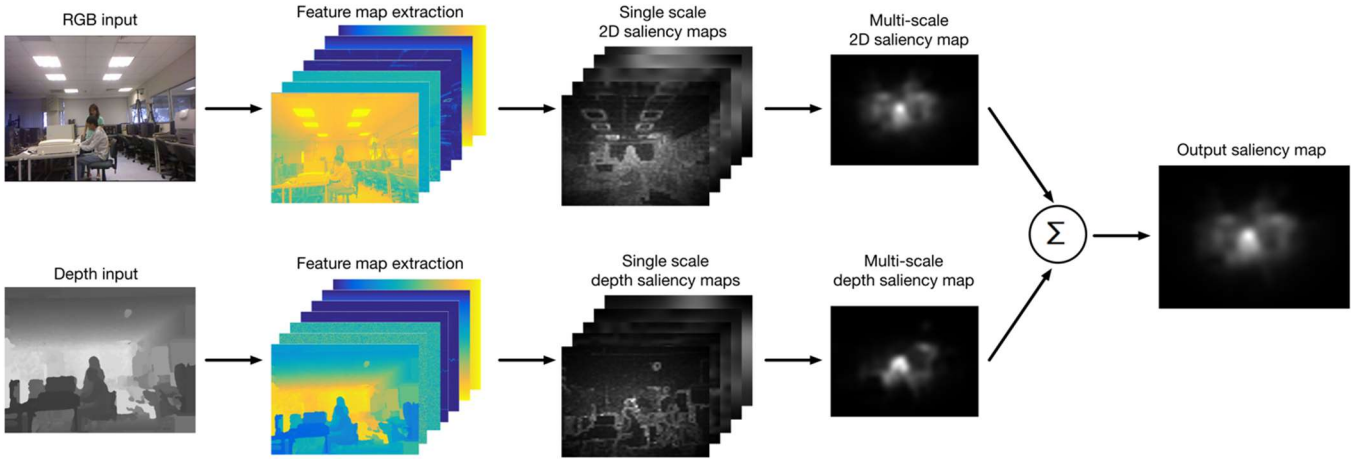

Figure 2. The proposed framework for depth-aware visual saliency estimation (see text for a detailed description).

The distances $d(R_i, R_j)$ in the second term measure the dissimilarity of the image regions R_i and R_j in terms of how different the corresponding feature statistics are, and correspond to the center-surround distances which are averaged over the m most similar image regions to R_i.

Erdem and Erdem proposed two different saliency models by considering two different distance definitions, one depends only on second order feature statistics. i.e. region covariances, (CovSal Model 1) and the second one depends on both first and second order feature statistics (CovSal Model 2).

In CovSal Model 1, the distances are estimated by considering the metric $\rho(\boldsymbol{C}_i, \boldsymbol{C}_j)$ proposed by [9] which is used to compare two covariance matrices $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$ representing the second order feature statistics collected from R_i and R_j, respectively:

$$d(R_i, R_j) = \rho(\boldsymbol{C}_i, \boldsymbol{C}_j) \,/\, \left(1 + \|x_i - x_j\|\right) \quad (4)$$

In *CovSal Model 2*, on the other hand, the distances are computed by considering the Sigma Points [13] representation $\Psi(\boldsymbol{C}_i, \boldsymbol{\mu}_i)$ which is formed by transforming the covariance matrices on Euclidean vector space using the Cholesky decomposition and enriching it with the mean of the feature vectors $\boldsymbol{\mu}_i$:

$$d(R_i, R_j) = \|\Psi(\boldsymbol{C}_i, \boldsymbol{\mu}_i) - \Psi(\boldsymbol{C}_j, \boldsymbol{\mu}_j)\| \,/\, \left(1 + \|x_i - x_j\|\right) \quad (5)$$

In both of the models, the distances are weighted by the inverse spatial distance between the regions, decreasing the influence of visually similar nearby regions.

In [8], five different saliency maps are extracted by changing the value of the scale parameter n, i.e., by varying the size of the regions where the statistics are collected. Then, a final multi-scale saliency map is computed by the pixelwise product of these individual single scale maps which are resized to the original image size. Finally, a Gaussian smoothing is applied to the estimated saliency maps.

3.2. Proposed Extension for Additional Depth Processing

In our study, we extend the CovSal models reviewed in the previous section to employ additional depth information available in RGB-D images. Since we process a RGB-D image, we assume that monocular depth cues have been extracted previously from the raw stereoscopic disparity information. Throughout the paper, we refer to these extensions as "Model1 + depth" or "Model2 + depth" according to their base saliency models. One of the key aims of this study is also compare and contrast the effectiveness of the CovSal saliency models with and without the extra depth information.

Figure 2 demonstrates the system architecture of the proposed depth-aware saliency models. Our models take a RGB-D image as an input, and produce a single saliency map extracted by considering the additional depth information. In our formulation, we first decompose this image into two as an RGB image and a depth image. While the RGB component have the information regarding the appearance of the scene, the depth component includes the depth information aligned with the appearance image as its name implies. In particular, we process each one of these components independently of the other by using the procedure discussed in the previous section, and then combine the extracted saliency maps for the final output.

The first step in our two stream framework includes extracting the visual features introduced in Equation (1) from the RGB and depth input images. Since these inputs capture different characteristics about the scene, the extracted features have different but complementary roles in saliency prediction. In the next step, these feature maps are used to collect first and second order region statistics which are then employed to extract single scale saliency maps. For a pixel, these single scale saliency maps point out how much the pixel is dissimilar to its surroundings in terms of appearance (for RGB stream) and depth (for depth stream). Then, these single scale saliency maps are combined by a pixel-wise multiplication operation to obtain two multi-scale saliency maps, one for 2D saliency S_{RGB} and the other for depth saliency S_{depth}.

Once these multi-scale saliency maps are extracted, as our final step, we integrate these maps into a final saliency maps by employing the following linear combination operation:

$$S = \lambda S_{RGB} + (1 - \lambda) S_{depth} \quad (6)$$

where the parameter λ denotes how much the final map depends on the extracted 2D and depth saliency maps. In our experiments, we found that setting $\lambda = 0.15$ gives accurate predictions.

4. Experiments

Performance of a saliency model is generally evaluated by comparing the resulting saliency map against the recorded eye

fixation data of human subjects collected for the given input image. In our study, we follow a similar strategy and test the performance of our models and the effect of including the depth information on the NUS-3D benchmark dataset [22]. In the following, we first describe this benchmark dataset and then review the metrics that are commonly used while evaluating the prediction of the human eye movements. After that, we analyze the outcome of our models on the NUS-3D dataset and present our qualitative and quantitative results.

4.1. Dataset

The NUS-3D dataset [22] is one of widely used datasets for depth-aware saliency estimation in the literature. It includes 600 indoor and outdoor natural RGB and aligned depth images, each captured around the National University of Singapore (NUS) campus via Microsoft Kinect and each having a resolution of 640 by 480 pixels. The eye fixation data were collected from 80 participants by following two different settings in free-viewing conditions. In the first setting, each participant views 100 3D images randomly selected from the set of 600 images on a 3D display by using active shutter glasses. In the second setting, each participant this time views another randomly selected 100 images in 2D display mode with the active shutter display switched off. Hence, the 2D and 3D eye fixation data are available for each image. Moreover, the researchers provide the raw and the smoothed depth maps extracted by the Microsoft Kinect sensor. In our experiments, we use the smoothed depth maps and carry out our experiments by employing both 2D and 3D ground truth fixation data.

4.2. Evaluation Metrics

For performance evaluation, we use four different metrics which are commonly used in saliency prediction. These are area under the ROC curve (AUC) and its shuffled version (sAUC), Pearson's correlation coefficient (CC), and the Kullback–Leibler divergence (KL-div). For all these evaluation metrics, we use the codes that are available online on the MIT Saliency Benchmark webpage (http://saliency.mit.edu) [6]. We report the mean values of these metrics averaged over all the images in the NUS-3D data set. These metrics compare the generated saliency map with the provided groundtruth fixation density maps and return real scalar values.

AUC and sAUC metrics treat the saliency estimation as a classification problem with the aim of classifying pixels as fixated or not-fixated. This is achieved by thresholding the given saliency map at specific saliency values, and by labeling the pixels above the threshold as fixated, and the remaining ones as not-fixated. Accordingly, for each threshold value, a true positive rate and a false positive rate are estimated by comparing the results with the ground truth eye fixation data. Then, the area under the corresponding receiver operator characteristics is the used as a value reflecting the classification performance of a saliency model. While the ideal AUC score is 1, the performance of random classification is around 0.5. Since, AUC might suffer from the tendency of humans to focus on the center location of an image, we also report sAUC scores which accounts for this center bias by collecting the negative samples not from the given image, but from the fixation points of randomly selected other images [32].

CC metric considers any saliency map of a given image and the corresponding ground truth fixation density map as two random variables and measures the strength of the linear association between them as a measure of similarity. A CC score of 1 indicates a perfect correlation, i.e. the saliency map is identical to the ground truth map. On the other hand, a score of 0 denotes that the generated saliency map and the ground truth are completely uncorrelated.

KL-Div metric is another distribution-based metric, which treats the generated saliency map and the ground truth fixation map as probability distributions, and measures the dissimilarity between these distributions. Hence, a KL-div value of 0 states that he generated saliency map and the ground truth fixation map are statistically the same, and the larger values of KL-Div indicate that they are significantly different from each other.

Results

We analyze the effectiveness of the proposed depth-aware saliency models on the NUS-3D dataset, which includes ground truth eye fixation data for both 2D and 3D images. Table 1 and Table 2 present the results of this quantitative analysis where we compared the generated saliency maps against the provided human fixation density maps. They show the evaluation scores of the generated saliency models, with and without depth information, along with another recently proposed depth-aware saliency model proposed by Hu et al. [15]. In particular, in Table 1, we use the 2D fixations collected by considering 2D viewing conditions, and in Table 2 we employ the 3D fixations data while evaluating the models. As mentioned above, the difference between these collected data lies in whether the active shutter glasses are switched off or on when participants explore a given image. While they are switched on, they view a 3D image and directly perceive the depth. On the hand, in the second case when they are switched off, they observe traditional 2D images.

As can be seen from Table 1 and Table 2, for both the 2D and 3D viewing settings employing additional depth information to predict saliency of an image in general improves the scores of all of the evaluation metrics except AUC. For the AUC metric, we observe a very slight decrease in the performance for the saliency models with depth information. As discussed in the previous section, AUC is a highly criticized evaluation metric for saliency for its inability to deal with the center bias [7], and we only report it due to its historical importance. For the CC and KL-div metrics, our first model which is the extension of the first CovSal model that considers only second order feature statistics gives better than our second model which additionally considers the first order statistics. On the other hand, for the sAUC metric, the relation is the opposite that our second model gives the better results. This clearly demonstrates the complementary nature of these evaluation metrics [7].

Moreover, the improvements obtained by the saliency models with the additional depth information can be regarded as relatively small, we also perform some statistical tests to determine whether the results with and without the additional depth information change in a statistically significant manner or not. For this, for each evaluation metric we carry out a two-tailed, paired Student t-test with a significance level of α=0.05 by employing the Benjamini-Hochberg correction [1] which is for multiple comparisons. For all the metrics, we report the adjusted p-values obtained by these tests in Table 1 and Table 2. It is important to note that the scores of the evaluation metrics with and without the additional depth information differ in a statistically significant manner. Only exception is the AUC metric that it does not pass the test for our first model when the 3D viewing condition is used. In our experiments, we also compare our results with that of the depth-aware proto-object based saliency model proposed by Hu et al. [15] As the results given in Table 1 and Table 2 indicate, the proposed saliency models gives more accurate predictions in terms of AUC and CC.

In Figure 3, we present sample images from the NUS-3D dataset (2D color and depth images) along with the ground truth fixation density maps and the saliency maps with and without the depth information generated by our first model. For the provided images, it is clearly visible that combining the standard 2D saliency maps with the generated depth saliency maps produces perceptually better maps against the ground truth fixations. For instance, in the first row, adding the depth saliency makes the tennis ball on the left visually more salient as compared to the original saliency map without depth information.

5. Conclusion

In this paper, we propose two new visual saliency models which employ additional depth information to predict where humans look at images. Our depth-aware models depend on the previously suggested CovSal models [8] which consider first and second order feature statistics to estimate the saliency. Specifically, we extend these models to process RGB-D images by extracting and combining 2D image saliency and depth saliency maps from color and depth images, respectively. Our analysis on the NUS-3D dataset with both 2D and 3D fixations indicate that including the depth information into these models gives more accurate saliency maps as compared to the saliency maps produced by the models that lack the depth information. These results point out that any saliency model can benefit from monocular depth cues, agreeing with the previous findings reported in [15,27]. As a final remark, it is important to note that how to integrate depth information into saliency prediction is an open problem that needs further work.

Table 1. Quantitative analysis of the proposed depth-aware saliency model on the NUS-3D datasets by considering the 2D fixations data. Employing depth information in general improves the accuracy of the saliency predictions in a statistically significant way

(verified by a paired t-test with a significance level of α=0.05 and with Benjamini-Hochberg correction).

	Evaluation metrics			
Saliency Model	AUC	sAUC	CC	KL-Div
Model1	0.8267	0.6308	0.4249	1.4992
Model1 + depth	**0.8260**	**0.6341**	**0.4280**	**1.4789**
p-value	4.27×10^{-2}	7.73×10^{-24}	2.07×10^{-5}	1.54×10^{-14}
Model2	0.8039	0.6566	0.3958	1.5065
Model2 + depth	**0.8023**	**0.6586**	**0.3924**	**1.5136**
p-value	1.98×10^{-92}	1.76×10^{-32}	3.14×10^{-31}	3.85×10^{-30}
Hu et al., 2016	0.7740	-	0.3590	1.485

Table 2. Quantitative analysis of the proposed depth-aware saliency model on the NUS-3D datasets by considering the 3D fixations data. Employing depth information in general improves the accuracy of the saliency predictions in a statistically significant way

(verified by a paired t-test with a significance level of α=0.05 and with Benjamini-Hochberg correction).

	Evaluation metrics			
Saliency Model	AUC	sAUC	CC	KL-Div
Model1	0.8286	0.6361	0.4019	1.5481
Model1 + depth	**0.8283**	**0.6398**	**0.4045**	**1.5249**
p-value	0.3630	5.78×10^{-25}	1.89×10^{-4}	5.55×10^{-17}
Model2	0.8044	0.6595	0.3664	1.5776
Model2 + depth	**0.8029**	**0.6615**	**0.3629**	**1.5853**
p-value	2.42×10^{-99}	3.43×10^{-33}	2.74×10^{-29}	1.90×10^{-31}
Hu et al., 2016	0.7770	-	0.3470	1.559

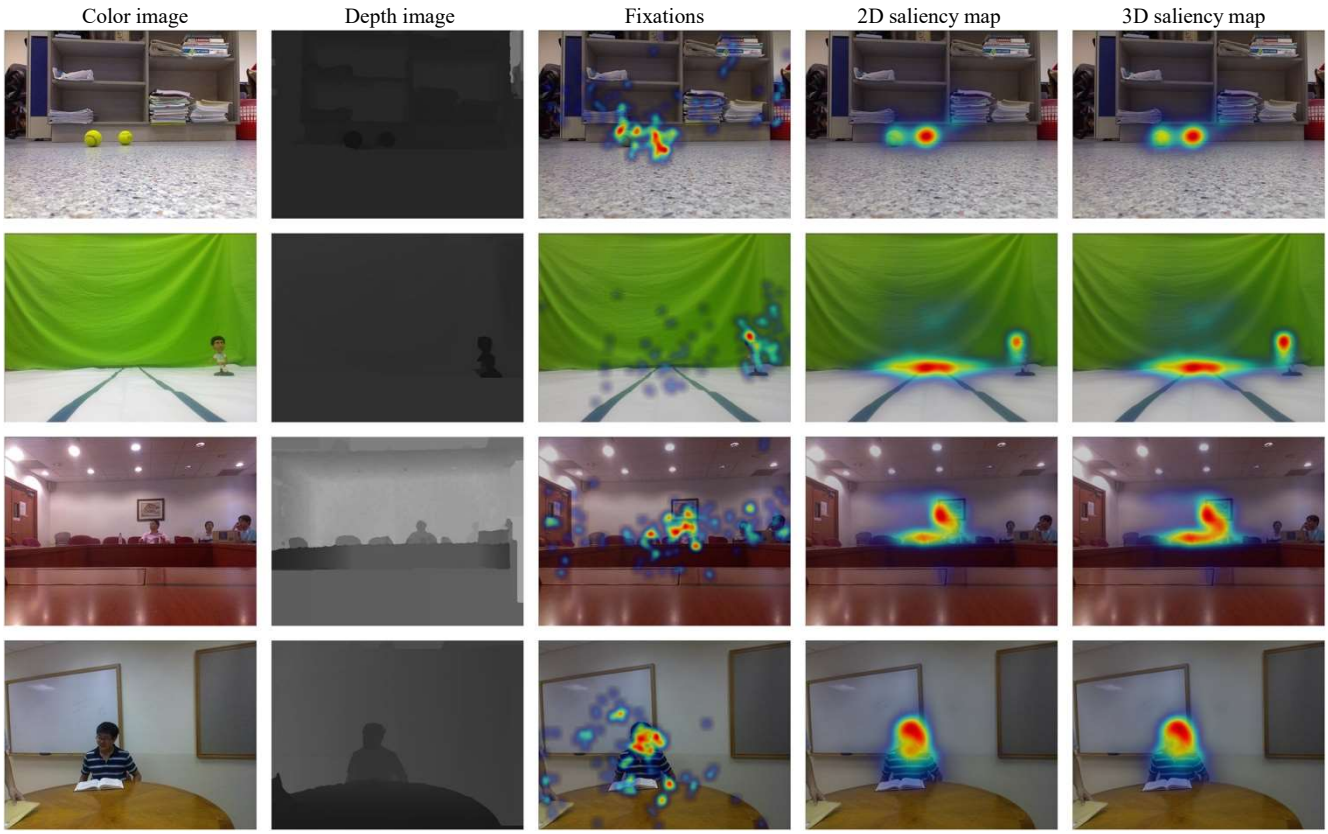

Figure 3. Sample qualitative results. Columns (left to right) show the color image along with the corresponding depth image, the ground truth fixations, and the generated 2D saliency map (without extra depth information) and the 3D saliency map (with extra depth information).

Acknowledgements

This research was supported in part by The Scientific and Technological Research Council of Turkey (TUBITAK), Career Development Award 112E146.

References

[1] Y. Benjamini, and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), pages 289-300.

[2] A. Borji, and L. Itti (2013). State-of-the-art in Visual Attention Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35(1), pages 185-207.

[3] N. D. Bruce, and J. K. Tsotsos (2005). An attentional framework for stereo vision. In Proc. IEEE Canadian Conference on Computer and Robot Vision, pages 88-95.

[4] N. Bruce, and J. Tsotsos (2006). Saliency based on information maximization. In Proc. Advance in Neural Information Processing Systems (NIPS), pages 155-162.

[5] N. Bruce, and J. Tsotsos (2009). Saliency, attention, and visual search: An information theoretic approach. Journal of Vision, Vol. 9(3):5, pages 1-24.

[6] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba (accessed by 2016). MIT Saliency Benchmark, http://saliency.mit.edu.

[7] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand (2016). What do different evaluation metrics tell us about saliency models?. arXiv preprint arXiv:1604.03605.

[8] E. Erdem, and A. Erdem (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. Journal of Vision, Vol. 13(4):1, pages 1-20.

[9] W. Föerstner, and B. Moonen (1999). A metric for covariance matrices (Tech. Rep.). Department of Geodesy and Geoinformatics, Stuttgart University, Germany.

[10] D. Gao, and N. Vasconcelos (2007). Bottom-up saliency is a discriminant process. In Proc. IEEE International Conference on Computer Vision (ICCV), pages 1-6.

[11] S. Goferman, L. Zelnik-Manor, and A. Tal (2010). Context-aware saliency detection. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2376-2383.

[12] J. Harel, C. Koch, and P. Perona (2007). Graph-based visual saliency. In Proc. Advance in Neural Information Processing Systems (NIPS), pages 545-552.

[13] X. Hong, H. Chang, S. Shan, X. Chen, and W. Gao (2009). Sigma Set: A small second order statistical region descriptor. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1802-1809.

[14] X. Hou, and L. Zhang (2007). Saliency detection: A spectral residual approach. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1-8.

[15] B. Hu, R. Kane-Jackson, and E. Niebur (2016). A proto-object based saliency model in three-dimensional space. Vision Research, Vol. 119, pages 42-49.

[16] H. Hügli, T. Jost, and N. Ouerhani (2005). Model performance for visual attention in real 3d color scenes. In Proc. Artificial intelligence and knowledge engineering applications: A bioinspired approach, pages 469-478

[17] I. Iatsun, M.-C. Larabin, C. Fernandez-Maloigne (2015). A visual attention model for stereoscopic 3D images using monocular cues. Signal Processing: Image Communication, Vol. 38, pages 70-83.

[18] L. Itti, C. Koch, and E. Niebur (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20(11), pages 1254-1259.

[19] T. Jost, N. Ouerhani, R. von Wartburg, R., Müri, and H.

Hügli (2004). Contribution of depth to visual attention: Comparison of a computer model and human. In Proc. Early Cognitive Vision Workshop, pages 1-4.

[20] T. Judd, K. Ehinger, F. Durand, and A. Torralba (2009). Learning to predict where humans look. In Proc. IEEE International Conference on Computer Vision (ICCV), pages 2106-2113.

[21] S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. V. Babu (2016). Saliency Unified: A Deep Architecture for Simultaneous Eye Fixation Prediction and Salient Object Segmentation. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5781-5790.

[22] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan (2012). Depth matters: Influence of depth cues on visual saliency. In Proc. European Conference of Computer Vision (ECCV), pages 101-115.

[23] C.-Y. Ma, and H.-M. Hang (2015). Learning-based saliency model with depth information. Journal of Vision, Vol. 15(6):19, pages. 1-22.

[24] R. Margolin, A. Tal, and L. Zelnik-Manor (2013). What makes a patch distinct? In Proc IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1139-1146.

[25] N. Ouerhani, and H. Hügli (2000). Computing visual attention from scene depth. In Proc. International Conference on Pattern Recognition, pages 375-378.

[26] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. O'Connor (2016). Shallow and deep convolutional networks for saliency prediction. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 598-606.

[27] S. Ramenahalli, and E. Niebur (2013). Computing 3D saliency from a 2D image. In Proc. Annual conference on information sciences and systems (CISS), pages 1-5.

[28] A. F. Russell, S. Mihalas, R. von der Heydt, E., Niebur, and R. Etienne-Cummings (2014). A model of proto-object based saliency. Vision Research, Vol. 94, pages 1-15.

[29] H. J. Seo, and P. Milanfar (2009). Static and space-time visual saliency detection by self-resemblance. Journal of Vision, Vol. 9(12):15, pages 1-27.

[30] O. Tuzel, F., Porikli, and P. Meer, (2006). Region covariance: A fast descriptor for detection and classification. In Proc. European Conference of Computer Vision (ECCV), pages 589-600.

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

[32] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell (2008). SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision, Vol. 8(7):32, pages 1-20.

[33] Y. Zhang, G. Jiang, M. Yu, and K. Chen (2010). Stereoscopic visual attention model for 3D video. In Proc. Advances in Multimedia Modeling, pages 314–324.