Depth really Matters: Improving Visual Salient Region Detection with Depth

Karthik Desingh[1]
http://researchweb.iiit.ac.in/~karthik.d/
K. Madhava Krishna[1]
http://www.iiit.ac.in/~mkrishna/
Deepu Rajan[2]
http://www.ntu.edu.sg/home/ASDRajan/
C.V. Jawahar[1]
http://www.iiit.ac.in/~jawahar/

[1] IIIT - Hyderabad
Hyderabad, India
[2] Nanyang Technological University
Singapore

Abstract

Depth information has been shown to affect identification of visually salient regions in images. In this paper, we investigate the role of depth in saliency detection in the presence of (i) competing saliencies due to appearance, (ii) depth-induced blur and (iii) centre-bias. Having established through experiments that depth continues to be a significant contributor to saliency in the presence of these cues, we propose a 3D-saliency formulation that takes into account structural features of objects in an indoor setting to identify regions at salient depth levels. Computed 3D saliency is used in conjunction with 2D saliency models through non-linear regression using SVM to improve saliency maps. Experiments on benchmark datasets containing depth information show that the proposed fusion of 3D saliency with 2D saliency models results in an average improvement in ROC scores of about 9% over state-of-the-art 2D saliency models.

1 Introduction and Related Work

Salient region detection has attracted much attention recently due to its ability to model the human visual attention mechanism, which has its roots in psychology but has been a topic of research in diverse areas such as neuroscience, robotics and computer vision. Identification of salient regions finds applications in object recognition [24], image retargeting [1], visual tracking [17] etc. There are two main approaches to salient region detection – top-down and bottom up, where the former is task dependent while the latter seeks to identify *pop-out* features that enable the extraction of distinct regions in an image. Bottom up saliency models have been developed as a pre-processing step to prioritize the search space for object detection tasks reducing the computational overhead [9]. Top-down approaches include [20] for scene recognition and [8] for tracking. Saliency detection has also been used as a pre-processing step for active segmentation of the objects in point clouds for manipulative tasks in robotics [3, 11].

Computational models have typically modelled saliency as a certain uniqueness or non repetitiveness of an area or pixel based on some features. For example Achanta *et al*. [2]

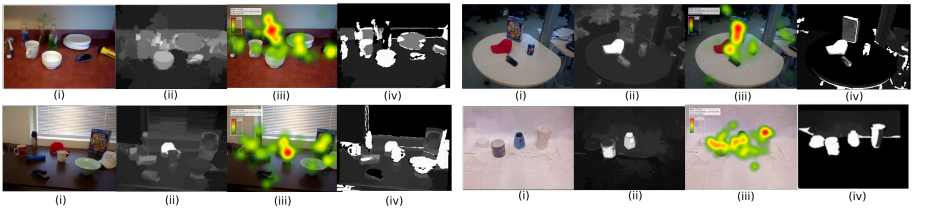


Figure 1: Four different scenes and their saliency maps; For each scene from top left (i) Original Image, (ii) RGB-Saliency map using RC [7], (iii) Human fixations from eye-tracker and (iv) Fused RGBD-saliency map

proposed a frequency tuned (FT) model that computes the pixel's saliency as a difference of the its color from the average image color. Zhai *et al*. [25] (LC) gives the saliency based on its contrast to all other pixels using only the luminance cue. Hou *et al*. [9] gave a spectral residual (SR) method that computes the saliency in the spectral domain. Cheng *et al*. [7] proposed two methods with one on the histogram based contrast (HC) and other on region based contrast (RC) and claim that the performance of RC is superior compared to FT, SR and LC.

With the advent of the Kinect sensor, depth information has been used in addition to color images for object recognition [13], human action recognition [22] and saliency detection [14, 16, 18]. In [18] and [16] , authors incorporate depth information from stereopsis making use of disparity maps of saliency detection. This implies that the accuracy of saliency maps (grayscale image showing salient regions with brighter intensities) depends on the disparity maps, which are not accurately obtained for cluttered indoor settings. Their work is limited to well framed images using stereoscopic cameras and does not cater to the needs of indoor environment. In [14], the authors use the Kinect sensor to obtain the depth and integrate it with a 2D saliency model. They develop a large 3D dataset along with fixations using a 3D eye-tracking system which is first of its kind. They study the spatial distribution of human fixations on 2D and 3D images and draw conclusions to the effect that incorporating depth information improves the quality of saliency maps. These conclusions effect in deciding priors that could be used to enhance the existing saliency maps.

Our work contrasts with [14] through the additional observations on depth saliency reported from our experiments, through the formulations of our 3D saliency model and the model for fusing 3D and visual saliency. It is well known that in images of large depth of field scenes taken using conventional cameras, the farther regions are out-of-focus, but images from the Kinect camera does not contain this phenomenon. Moreover, there is a bias towards the centre of the image by the human visual system during fixation [23]. Our experiments on depth-induced blurred images and on the centre-bias characteristic further reinforce the importance of depth in visual saliency. We also conduct experiments to study the role of depth in saliency detection when there are competing saliencies attributed to appearance, such as color contrast (this was also not done in [14]).

The main contributions of this paper are: (i) The development of a 3D saliency model that integrates depth and geometric features of object surfaces in indoor scenes (ii) Fusion of appearance (RGB) saliency with depth saliency through non-linear regression using SVM (iii) Experiments to support the hypothesis that depth improves saliency detection in the presence of blur and centre-bias. The effectiveness of the 3D-saliency model and its fusion with RGB-saliency is illustrated through experiments on two benchmark datasets that contain

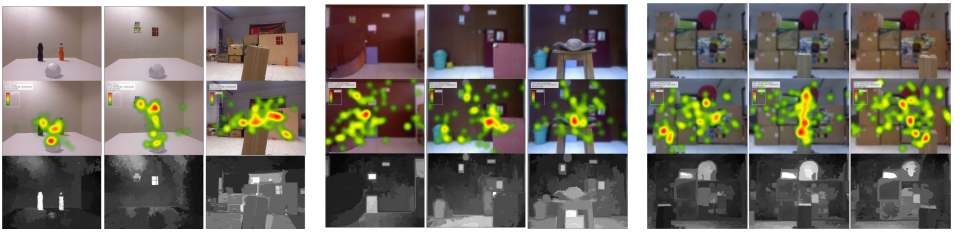
depth information – University of Washington RGB-D dataset [12] and Berkely 3D object dataset [10]. Current state-of-the-art saliency detection algorithms perform poorly on these datasets that depict indoor scenes due to the presence of competing saliencies in the form of color contrast. For example in Fig. 1, saliency maps of [7] is shown for three different scenes, along with its human eye fixations and our proposed saliency map after fusion. It is seen from the top left scene of Fig. 1, that illumination plays spoiler role in RGB-saliency map. In bottom left scene of Fig. 1, the RGB-saliency is focused on the cap though multiple salient objects are present in the scene. Bottom right scene of Fig. 1, shows the limitation of the RGB-saliency when the object is similar in appearance with the background.

2 Effect of Depth on Saliency

The correlation and influence of depth cues in modelling saliency was studied in [14]. Based on fixations on 2D and 3D images, they conclude that humans fixate preferentially at closer depth ranges. They determine that the relation between depth and saliency is non-linear. However, they do not consider three important issues in their study. Firstly, as mentioned in the previous section, what is the effect of depth on saliency in the presence of competing saliencies in the background? In other words, if there is a high color contrasting object in the background, will the foreground object closer to the camera still capture saliency? Secondly, a conventional camera looking at a large depth-of-field scene will be focused at one depth implying that objects lying at other depths will be blurred. In such a situation, blur adds to the effect of depth in determining the salient regions. The third issue is that of centre-bias which implies that human fixations are biased to the centre of the screen when viewing 2D data. Would such a bias exist even when viewing large depth-of-field images? In this section, we answer these questions through experiments on each of the three cases with 15 images for each case and analyze human fixations on them. Eight participants (4 male and 4 female) were shown the images. Images were displayed for 6 seconds. The observations are as follows.

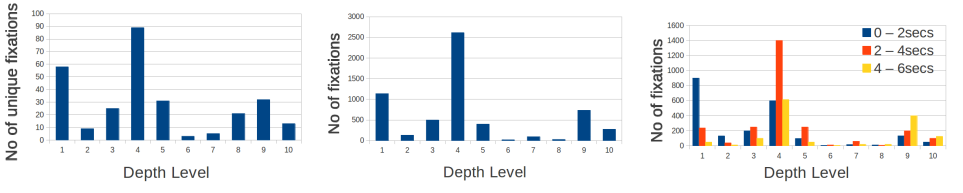
Competing saliency: Typical indoor settings have been created and captured by Kinect depth camera, which, it must be noted, does not have option to focus at a depth. It can be seen from Fig. 2(a) that objects lying closer to the camera and whose appearance does not contrast with the background are fixated by human subjects and these fixations are comparable to the other regions in the scene. However the RC saliency model is not able to capture this information, as shown in the last row, since it considers only appearance.

Depth levels are the quantized levels of the depth range of the particular set of images. In all these images the object closer to the camera is placed at a distance of 0.5 meters. Hence the depth level 1 is the one that constitutes to the fixations on the bland object in the experiment. Fixations are analyzed at each depth levels and plotted as unique fixations, repetitive fixations and temporal fixations. Fig. 3(a) shows that the low contrast object at closer depth gets equivalent unique fixations when compared to farther attentive regions. However the closer objects are not fixated for large period of time to get multiple fixations, as seen from Fig. 3(b). Hence we see darker red spots are on objects that are not closest to the camera in the Fig. 2(a). Observing the sequence of fixations by Fig. 3(c), we note that low contrast object at closer depth gets more attention in initial couple of seconds than the later stage, which attests to the temporal characteristics of visual attention.



(a) Competing saliency (b) Blurred (c) Centre-bias

Figure 2: (Top) Original images, (Middle) Human eye-fixations shown as a heat map on the count of fixations, (Bottom) Saliency map given by state-of-the-art model RC [7]. In all these settings objects at closer depth get comparable fixations, which is not reflected in the saliency model without depth cues



(a) (b) (c)

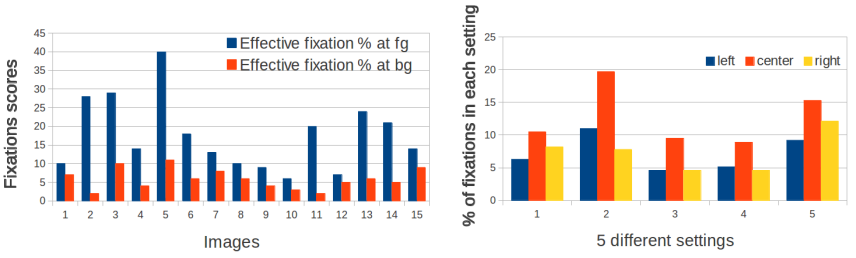
Figure 3: Depth stimulates the human fixations in practical indoor scenes; a) unique fixations vs depth levels, b) repeating fixations vs depth levels, c) temporal fixations vs depth levels

Blurred scenes: Fig. 2(b) shows scenes where the background has been blurred (since Kinect does not provide depth-induced blurred images) using the relationship between depth map and image captured by the Kinect. Image regions beyond a depth is blurred by Gaussian blur function of OpenCV [5] using Kernel size of (27, 27) with auto-computation of the sigma values enabled.

To know quantitatively how the fixations are at foreground and background, effective fixations at these levels are analyzed. Effective fixation at a foreground is the number of fixations per-pixel in the foreground region. Similarly the effective fixations at the background is computed and plotted for 15 images as shown in the Fig. 4(a). From this plot it is observed that effective fixations at the foreground is higher compared to the effective fixations at the background (blurred). This leads to an observations that the humans fixate on objects that are focused irrespective of whether the objects have low contrast with respect to the surroundings or not.

Centre-bias: In this experiment, when the foreground objects are placed left, center and right in the field of view, their fixations vary largely. Five sets of scenes with these 3 variations were setup to confirm this observation. One such set is shown in the Fig. 2(c) with the foreground object at left, center and right positions. Percentage of the foreground fixations at these 3 spatial locations for 5 settings are computed and plotted in the Fig. 4(b). This plot shows that the low-contrast object placed at the centre of the field of view gets more attention compared to other locations. Thus, the notion of centre bias is also applicable in large depth-of-field scenes.

We have shown that depth continues to affect saliency even when there are other attentive



(a) (b)

Figure 4: a) Percentage of Effective fixations at foreground and background for 15 images, b) Percentage of fixations for spatial locations left, right and center for 5 different settings

cues present in the image. In the next section, we develop a model to capture depth saliency in an indoor environment.

3 3D-saliency for Indoor Environment

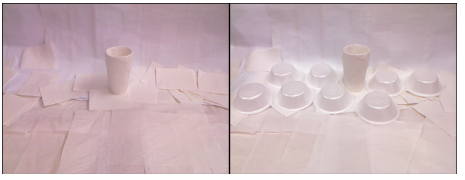
Having established through experiments that depth has an important role in identifying salient regions, we develop a method to measure saliency from depth information and the structural features of objects in the scene. We call this as 3D-saliency denoted by D. A challenging scenario in obtaining 3D-saliency is shown in Fig. 5(a), where there is very low contrast between the salient region – the tall mug – and the surroundings causing appearance based saliency techniques to fail. In such a situation, it is imperative to depend on depth to determine saliency.

Compared to the stereo technology, active projection approach used in depth sensors like Kinect results in reliable depth readings. The point cloud created from the depth image is segmented using a region growing technique [19] which is implemented in Point cloud library [21]. Features used in this region growing technique are curvature and smoothness of the surface. We adapt the region based contrast method from Cheng *et al.* [7] in computing contrast strengths for the segmented 3D surfaces/regions. Each segmented region is assigned a contrast score using surface normals as the feature. Structure of the surface can be described based on the distribution of normals in the region. We compute a histogram of angular distances formed by every pair of normals in the region. Every region R_k is associated with a histogram H_k. Contrast score C_k of a region R_k is computed as the sum of the dot products of its histogram with histograms of other regions in the scene. Since the depth of the region is influencing the visual attention, the contrast score is scaled by a value Z_k, which is the depth of the region R_k from the sensor. Z_k of the any region from the sensor is computed by finding the depth of the centroid region. Hence the constrast score becomes

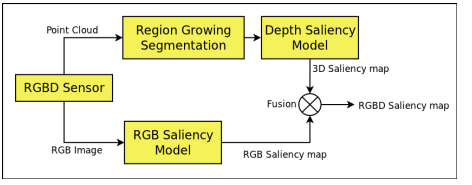
$$C_k = Z_k \sum_{j \neq k} D_{kj} \tag{1}$$

where D_{kj} is the dot product between histograms H_k and H_j.

Dimension of the regions after segmentation, plays a significant role in deciding the saliency of the region. Suppose there are only two regions in the scene whose surfaces are contrasting with each other. Since the contrast score calculated in the above section depends



(a)



(b)

Figure 5: a) Two images in which the salient objects are unique due to their geometric characteristics, b) Block diagram of proposed RGBD-saliency computation: Point cloud is segmented into regions, which are assigned saliency scores by Depth saliency module. The obtained 3D-saliency map is fused with the RGB-saliency map to obtain RGBD-saliency map

on the surface description through histograms, they get equal scores. However, in order to define the saliency, sizes of the regions i.e. the number of points in the region, have to be considered. We find the ratio of the region dimension to the half of the scene dimension. Considering n_k as the number of 3D points in the region R_k, Eq. 1 becomes

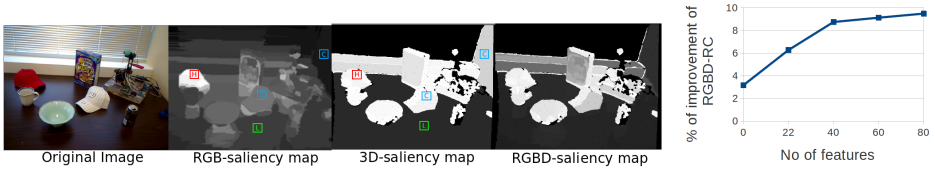
$$C_k = \frac{2Z_k n_k \sum_{j \neq k} D_{kj}}{\sum_j n_j} \tag{2}$$

The region with less C score is considered to be the one that is unique in the scene with respect to depth only. Hence saliency of the region R_k becomes $S_k = 1 - C_k/C_{max}$, where C_{max} is the maximum contrast score in the scene for a region. Having a one to one correspondence between every 3D point in the point cloud to a pixel in the image, the 3D-saliency map can be computed by assigning the saliency score to its corresponding pixel. With the obtained 3D-saliency map, we fuse saliency maps given by the state-of-the-art algorithms to obtain the RGBD-saliency map.

4 RGBD-Saliency Fusion

In this section, we describe a method to fuse depth (3D) saliency with 2D saliency models to obtain the final saliency map, which we call the RGBD-saliency map. Fig. 5(b) shows the block diagram of proposed fusion of depth and RGB-saliency, where the 3D-saliency is obtained for each region generated by a region segmentation of the point cloud.

Consider $S_{rgb}(x,y)$ as RGB-saliency and $S_{3D}(x,y)$ as 3D-saliency value for a pixel at (x,y) for below discussion. Both S_{rgb} and S_{3D} are high at the regions which are attentive in appearance and 3D shape marked as H in the Fig. 6(a). These cases can be obviously considered to be highly salient in the fusion. Similarly when both S_{rgb} and S_{3D} are low marked as L in the Fig. 6(a), they have to be considered as less salient in the fusion. Then there are the cases where S_{rgb} and S_{3D} conflict with each other. These complementary scenarios where one is high and the other is low are marked as C in Fig. 6(a). The fusion of such cases is not always straightforward for a high in one model and low in the other could be due to false positives making one of the saliency values high. Tricky are also those cases where S_{rgb} and S_{3D} depict average values.



(a) (b)

Figure 6: a) Fusion cases: Original image (first from left), RGB-saliency map(second), 3D-saliency map(third) and Fused RGBD-saliency map. Regions are marked H-high, L-low and C-compliment to show scores of both RGB-saliency and 3D-saliency and the regions where they compliment each other, b) Percentage improvement in fusion vs increase in the features

To generalize these cases we consider training SVM regression using libsvm [6] with some images and learn how to fuse the saliency maps. To avoid the computational expense we sample entire pixels of training images into training and validation data. Using libsvm we cross validate the same for varying C and γ of the SVM kernel. We choose the training model with least mean squared error in the validation. With this trained model, we fuse the values of RGB-saliency and 3D-saliency of test images and get the predicted value for each pixel. We also experimented with additional local features of the regions to improve the performance of the fusion process. Overall fusion by learning is a function of saliency scores, features and weights assigned by these features to determine its fused saliency score. This function is given by

$$rgbd_i = f(w, f_i, rgb_i, d_i) \tag{3}$$

where w is the weight vector learnt by the SVM model with the help of local feature vector f_i and saliency scores rgb_i, d_i to determine $rgbd_i$ at i^{th} pixel of an image.

Additional features used in the fusion process are (along with their feature lengths): *Color Histogram (30)* of region both in terms of RGB and HSV each of 15 bins . *Contour Compactness (1)* is the ratio of the perimeter to the area of the region. *Dimensionality (2)* is the two ratios, minimum dimension by maximum dimension and medium dimension by maximum dimension. *Perspective score (8)* is the ratio of the area projected in the image to the maximum area spread by the region in 3D. *Discontinuities with neighbours (10)* is measure of how much the region is connected with its neighbouring regions. *Size and Location (9)* of the region with respect to the scene gives the range and location of the region in three dimension. *Location* here constitutes to the scaled location of the region with respect to the scene by computing min and max values in each dimension. This takes into account of our third observation of spatial context in the Section 2. *Verticality (20)* is the histogram measure of difference of the normals in the region with respect to the camera pose. They combinely constitute a feature length of 82 along with the RGB and 3D-saliency score. Fig. 6(b) shows the improvement in the performance with the addition of these features.

5 Experiments

We start by discussing the dataset and the benchmarking techniques. By fusing the proposed 3D-saliency with the available RGB-saliency models we show significant improvement in ROC scores of the generated saliency maps.

Table 1: ROC scores of saliency models on UW dataset images

Saliency Models	RGB	D	RGB-D	% change in RGB
FT	0.6433	0.7558	0.7975	↑ 15.42
LC	0.5748	0.7558	0.7994	↑ 22.46
HC	0.5980	0.7558	0.7912	↑ 19.32
SR	0.7838	0.7558	0.8347	↑ 5.09
RC	0.7105	0.7558	0.8053	↑ 9.48

Table 2: ROC scores of saliency model RC [7] for subset of Our dataset images categorized for blurred and spatial variations in the experiments

Category of Images	RGB-RC	D	RGBD-RC	% improvement
Our dataset - Blurred	0.6881	0.7016	0.7391	↑ 5.10%
Our dataset - Spatial variations	0.7688	0.7138	0.8267	↑ 5.79%

Datasets and Benchmarking: Public benchmark datasets for evaluating saliency algorithms available, include only monocular images without depth maps. To the best of our knowledge, there is no publicly available RGB-D database for saliency analysis. In order to test the RGBD-saliency and make comparisons, we used RGB-D dataset provided by the University of Washington (UW) [12] and also the Berkeley 3D object dataset [10]. In addition to this we generate our own dataset with 33 images. These datasets have different scene categories, out of which we choose 28 images from UW dataset and 50 images from Berkeley dataset, which are distinct in terms of back ground and objects for our experiments. In the fusion process with UW dataset, we train on 4 images and test on 24. With Berkeley dataset we train on 10 images and test on 40. Similarly we train on 6 images and test on 27 in our own dataset captured using Kinect sensor.

We create ground truth by region based method [15]. Eight subjects with 4 males and 4 females of non-technical background were requested to draw bounding borders around objects/regions (maximum number of objects allowed in marking is 4) that attracts them in the image. It is noticed that the objects marked by the subjects under the scenes where there are many objects in the scene had inconsistency in being a salient ground truth. Hence we set the pixel value to 1 if at least 2 subjects agree that the pixel it belongs to a salient region and zero otherwise.

Performance Evaluation and Results: Experiments are performed to show how the proposed RGBD-saliency enhances the performance of existing saliency models, across different datasets. ROC scores showing the improvement in performance of five RGB-saliency models after fusing with the 3D-saliency is shown in Table 1 for UW dataset. It can be observed that the fusion improves the scores across all the five models by a significant amount. Scores showing the performance of the fused RGBD-saliency for (RC [7]), across three datasets is shown in Table 3. This concludes that the improvement is not specific to a particular setting. ROC scores are computed for our dataset where experimental settings with blurred images and spatial variations are categorized. Improvements in scores is shown in Table 2.

UW dataset is entirely lab/workspace setting, whereas Berkeley 3D dataset also includes household settings along with lab/workspace scenes. Our own dataset is taken at settings as discussed in Section 2 for experimental analysis which also includes indoor settings that is entirely different from the other two datasets. All these three datasets are different from

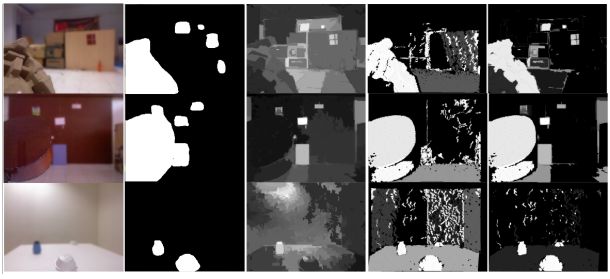


Figure 7: From left (1) Original image and its (2) Human annotated ground truth, (3) RGB-saliency map using RC [7], (4) 3D-saliency map and the (5) Fused RGBD-saliency map. RGB-saliency fails to map the objects that are closer, but the fusion of 3D-saliency helps in recovering these objects.

Table 3: ROC scores of saliency model RC [7] for all three datasets used in this work

Datasets	RGB-RC	D	RGBD-RC	% improvement
Univ of Washington	0.7105	0.7558	0.8053	↑ 9.48%
Berkeley 3D dataset	0.7246	0.7518	0.8157	↑ 9.11%
Our dataset	0.7287	0.7312	0.8001	↑ 7.14%

each other in their scenic structure and objects included. Hence it is worth evaluating the performance of the proposed saliency model on these datasets. Table 3 shows that RGB-saliency across all these datasets perform to a similar level, while the 3D-saliency performs superior compared to their visual saliency models. This superior performance of the 3D-saliency is because of the largely varying depth levels and structures in the indoor scenes. But however this alone does not constitute to a better saliency because, appearance is the primary cue to the visual attention. Hence the fusion is performed and it can be seen in Table 3, that fused RGBD model provides an improvement of around 9% across all these datasets. Having shown the improvement on the state-of-the-art method (RC) [7] in Table 1 we show the improvements across the other visual saliency models on UW dataset. Results of these models on other datasets is shown in supplementary material. Table 2 shows how the ROC scores are for the blurred and center biased setup in the Section 2. It can be infered that RGB-RC score of the Blurred category is less compared to the D and there is an improvement of 5.10% by the fusion process. Whereas in spatial variations, the images contain subset of images where the object closer to camera is placed at left and right locations, which decreases the D score compared to RGB-RC score but overall, improves the score by 5.79% when fused.

It can be seen from the Fig. 7, that the RGB-saliency fails to map the objects that are less contrast with the background. However fusion of 3D-saliency and RGB-saliency helps in recovering the objects that were missed out. It should also be noticed that the regions of the background which has slightly higher saliency score in pure 3D-saliency has been brought down to least score in the fusion. Hence both RGB-saliency and 3D-saliency compliment each other in the fusion process. Saliency maps of all the models discussed and proposed in this paper are shown in supplementary material across three dataset images.

6 Conclusion

In this work we proposed RGBD-saliency to resolve the drawbacks of the existing visual saliency models in a pratical indoor settings. We derived RGBD-saliency by formulating a 3D-saliency model based on region contrast of the scene and fused it with the existing saliency models using SVM. It is shown that the resulting fused model clearly outperforms the individual models by a significant amount of 9% on average. We test this behaviour successfully across different datasets and quantify the enhancements.

Acknowledgements

We would like to thank the Dept of Information Technology to have funded this work through the grants made available by the National Program on Perception Engineering - Phase 2

References

[1] R. Achanta and S. Susstrunk. Saliency Detection for Content-aware Image Resizing. In *ICIP*, 2009.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-Tuned Salient Region Detection. In *CVPR*, 2009.

[3] M. Bjorkman and D. Kragic. Active 3D Scene Segmentation and Detection of unknown Objects. In *ICRA*, 2010.

[4] A. Borji and L. Itti. State-of-the-art in Visual Attention Modeling. *TPAMI*, 2012.

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[6] C.C. Chang and C.J. Lin. LIBSVM: A Library for Support Vector Machines. In *TIST*, 2011.

[7] M.M. Cheng, G.X. Zhang, N.J. Mitra, X. Huang, and S.M. Hu. Global Contrast based Salient Region Detection. In *CVPR*, 2011.

[8] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Meissner, G. Bradski, P. Baumstarck, S. Chung, and A.Y. Ng. Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video. In *IJCAI*, 2007.

[9] X. Hou and L. Zhang. Saliency Detection: A Spectral Residual Approach. In *CVPR*, 2007.

[10] A. Janoch, S. Karayev, Y. Jia, J.T. Barron, M. Fritz, K. Saenko, and T. Darrell. A Category-Level 3D Object Dataset: Putting the Kinect to Work. In *ICCV Workshop*. 2011.

[11] M. Johnson-Roberson, J. Bohg, M. Björkman, and D. Kragic. Attention Based Active 3D Point Cloud Segmentation. In *IROS*, 2010.

[12] K. Lai, L. Bo, X. Ren, and D. Fox. A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. In *ICRA*, 2011.

[13] K. Lai, L. Bo, X. Ren, and D. Fox. A Scalable Tree-based Approach for Joint Object and Pose Recognition. In *AAAI*, 2011.

[14] C. Lang, T.V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan. Depth matters: Influence of Depth Cues on Visual Saliency. In *ECCV*. 2012.

[15] J. Li, M.D. Levine, X. An, X. Xu, and H. He. Visual Saliency Based on Scale-Space Analysis in the Frequency Domain. In *TPAMI*, 2012.

[16] F. Liu, X. Li, Y. Geng, and Y. Niu. Leveraging Stereopsis for Saliency Analysis. In *CVPR*, 2012.

[17] V. Mahadevan and N. Vasconcelos. On the Connections between Saliency and Tracking. In *NIPS*, 2012.

[18] A. Maki, P. Nordlund, and J.O. Eklundh. Attentional Scene Segmentation: Integrating Depth and Motion. In *CVIU*, 2000.

[19] T. Rabbani, F. van Den Heuvel, and G. Vosselmann. Segmentation of Point Clouds using Smoothness Constraint. In *ISPRS*, 2006.

[20] M. Rudinac and P.P. Jonker. Saliency Detection and Object Localization in Indoor Environments. In *ICPR*, 2010.

[21] R.B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). In *ICRA*, 2011.

[22] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human Activity Detection from RGBD Images. In *AAAI Workshop on Pattern, Activity and Intent Recognition*, 2011.

[23] P.H. Tseng, R. Carmi, I.G.M. Cameron, D.P. Munoz, and L. Itti. Quantifying Center Bias of Observers in Free Viewing of Dynamic Natural Scenes. *Journal of vision*, 2009.

[24] D. Walther, U. Rutishauser, C. Koch, and P. Perona. On the Usefulness of Attention for Object Recognition. In *ECCV Workshop on Attention and Performance in Computational Vision*, 2004.

[25] Y. Zhai and M. Shah. Visual Attention Detection in Video Sequences using Spatiotemporal Cues. In *ACM International conference on Multimedia*, 2006.