Gaze Location Prediction with Depth Features as Auxiliary Information

Redwan Abdo A. Mohammed, Lars Schwabe, and Oliver Staadt

University of Rostock, Institute of Computer Science, Rostock, Germany
{redwan.mohammed,lars.schwabe,oliver.staadt}@uni-rostock.de

Abstract. We present the results of a first experimental study to improve the computation of saliency maps, by using luminance and depth images features. More specifically, we have recorded the center of gaze of users when they were viewing natural scenes. We used machine learning techniques to train a bottom-up, top-down model of saliency based on 2D and depth features/cues. We found that models trained on Itti & Koch and depth features combined outperform models trained on other individual features (i.e. only Gabor filter responses or only depth features), or trained on combination of these features. As a consequence, depth features combined with Itti & Koch features improve the prediction of gaze locations. This first characterization of using joint luminance and depth features is an important step towards developing models of eye movements, which operate well under natural conditions such as those encountered in HCI settings.

1 Introduction

Being able to predict gaze locations, as compared to only measuring them, is desirable in many application scenarios such as video compression, the design of web pages and commercials adaptive user interfaces, interactive visualization, or attention management systems[14,5]. However, eye movements are known to depend on task demands, in other words, information not present in the visual stimulus. As a consequence, algorithms based on the computation of salient locations from only the bottom-up visual signals have principled limitations in gaze prediction.

Eye movements have been predicted mainly using purely stimulus-driven models. Most models of saliency [6,8] are biologically inspired and based on a bottom-up computational model which does *not* take into account contextual factors or the goal of a user in a visual task. Multiple low-level visual features such as intensity, color, orientation, texture and motion are extracted from the image at multiple scales. Then, a saliency map is computed for each of the features and combined in a linear or non-linear fashion into a master saliency map that represents the saliency of each pixel. This idea of saliency maps was used in other studies, where it was extended and further developed. For example, Mahadevan and Vasconcelos [3] proposed a discriminant formulation of center-surround

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 281–292, 2014.

saliency for static images. One can view their work as a normative approach, because they first formulate the saliency map computation as a problem, and then derive their algorithm as the solution to this problem. More specifically, they consider saliency as a decision making task informed by natural image statistics. The outcome of their work is an automatic selection of the important features. This improves the original Itti & Koch model [6], where the features selection and combination was done in a heuristic way. This was later also extended to dynamic scenes and movies using dynamic textures [8]. However, the original Itti & Koch model was also improved recently using graphs to compute saliency [4]. This shows that the concept of saliency maps is still very fruitful and can guide research in predicting eye movements. These saliency-based models are all based on low-level image features. Despite this limitation, they often predict gaze well, but mid- and high-level features also affect gaze. Therefore, Judd et al. [17] pursued a machine learning approach: They learned gaze points based on measured eye movements using a linear SVM and low-, mid- and high-level features. They reported better predictions than Itti & Koch on 1003 images observed by 15 subjects [17].

Another line of research has investigated the depth structure of natural scenes using range sensors [12,18,10]. This depth structure is not directly accessible to the human vision system and needs to be inferred using stereo vision or other depth cues. Some statistical aspects of depth images as well as the relation between depth and luminance images have been investigated before [18,10,13], but the statistical properties of depth images at the center of gaze are not clear [11]. For example, simple questions such as "Do humans look more often to high contrast edges due to depth gaps than to edges due to texture borders?" have not been addressed yet [11]. It was shown, however, that eye movements are far from a random sampling. It was even suggested that the statistics of natural images differ at the center of gaze when compared to random sampling [13]. Thus, taking into account eye movements is essential for shaping artificial vision systems via natural images. In [9] we have analyzed the saliency in 2D pixel and depth images using a very simple feature: the local standard deviation of pixels. We found that saliency in depth images is bimodally distributed with highly salient locations corresponding to low salient 2D image locations. Given that most saliency algorithms work on the 2D images, this finding points towards including depth cues into the computation of saliency maps.

In this paper, we present the results of a first experimental study to further improve the computation of saliency maps. More specifically, We have recorded the center of gaze of users when they were viewing natural scenes. We first examined the statistical characterization of depth features in natural scenes at the center of gaze. The rational for investigating depth images is that they may reveal the "saliency that matters", because when interacting with the environment we evolved by interacting with objects in a three dimensional (3D) world. Thus, we hypothesize that saliency maps respecting this will ultimately outperform saliency maps computed only on the basis of 2D pixel images in terms of predicting eye movements. We then examined the presence of depth features around

gaze locations. We used machine learning to train a bottom-up, top-down model of saliency based on 2D and depth features/cues. We used different performances distance measures. We found that models trained on Itti & Koch and depth features combined outperforms models trained on other individual features or other pairs of features combined.

This paper is organized as follows: First, we describe the material and methods including the image material (Sec. 2.1) and the features we extracted from the luminance and depth images (Sec. 2.3 and 2.4). Then, we present the results of our analysis, where we first compared the distribution of depth values of patches in the center of gaze to that expected from random sampling (Sec. 3) and then gaze location prediction when viewing photos of natural scenes (Sec. 4).

2 Material and Methods

2.1 Stimulus Material

Forty images obtained originally from Make3D project Range and Image Dataset [15,16] were presented to five subjects. The 2D color pixel images were recorded with a resolution of 1704×2272 pixels, but the depth images with a resolution of 305×55 pixels. They where 40 images from "forest scene", "city scene", and "landscape scene". The users were males and females between the ages of 18 and 35. Three of the viewers were researchers in institute of computer science and the others were naive viewers. All viewers sit at a distance of approximately *1.5 m* from the computer screen of resolution 1280x1024 in a dark room and used a chin rest combined with a bite bar to stabilize their head. An mobile eye tracker recorded their gaze path on a separate computer as they viewed each image at full resolution for ten seconds separated by two seconds of viewing a gray screen.

2.2 Measuring Gaze Locations

An iView X HED 4 Eye Tracking System (SMI) was used to record eye position. The eye tracker uses two cameras. The first is used to track the pupil and the second camera records the scene view. The gaze position is reported with a sampling rate of 50 Hz and a reported accuracy of 0.5°-1°. We used the default lens (3.6 mm) for the scene camera which provides a viewing angle of ±31° horizontally and ±22° vertically. The scene camera resolution is 752×480. Then, to avoid parallax error, we calibrated in a distance within 1-1.5 m. We used a calibration with five points so that the SMI recording software can compute the gaze location in scene camera coordinates from the recorded pupil images. The scene camera of the eye tracker delivers RGB frames as well as gaze locations, both with time stamps (Figure 1 a), Also we recorded information about which and when each image have been presented to the viewer. Our analysis were all done offline. First we aligned the frames temporally to the high resolution images using the information we recorded about when each image have been presented to the viewer. Then we used normalized Cross-Correlation [7] to register each

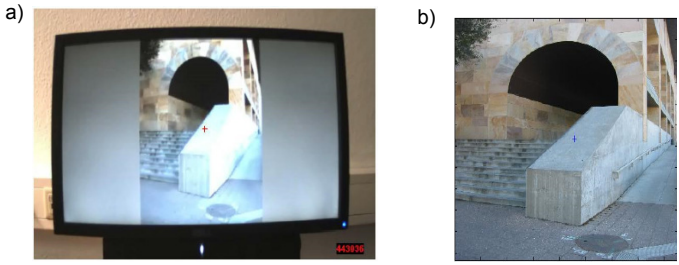

Fig. 1. Example of a gaze registration. **a)** Frame from the scene camera of the eye tracker and the corresponding gaze point (Red cross). **b)** Registered gaze point (Blue cross) on the corresponding high resolution image.

part of interest in each frame to the corresponding high resolution image. Using the transformation obtained to register each gaze point to the high resolution image (Figure 1 b), we generated a saliency map of the locations fixated by each viewer. Also, we convolve a Gaussian filter across the user's fixation locations in order to obtain a continuous saliency map of an image from the eye tracking data of a user.

2.3 Features of Luminance Images

Different low-level features were collected. For example: the intensity, orientation and color contrast channels as calculated by Itti and Koch's saliency method [6]. Also, each gray-scale image is linearly decomposed into a set of edge feature responses to Gabor filters with different orientations. We used orientations $\theta = \{0°, 15°, \ldots, 165°\}$, but only one frequency and two spatial phases. Within each image we subtracted the mean from the filter responses to each orientation, and normalized the responses to the interval between -1 and 1. We used Gabor filters responses to compare the performance with the 3D edges.

2.4 Features of Depth Images

Gap Discontinuity. A gap discontinuity in the underlying 3D structure is a significant depth difference in a small neighborhood. We measure gap discontinuity μ_{GD} by computing the maximum difference in depth between the depth of a pixel in the depth image and at its eight neighboring pixel. Here, we considered the methods presented in [19]; μ_{GD} for a point (x, y) is defined as:

$$\mu_{GD}(x, y) = \max\{ \mid z(x, y) - z(x + i, y + j) \mid : -1 \leq i, j \leq 1\}, \qquad (1)$$

where $z(x, y)$ represents a depth value. This quantity is then thresholded to generate a binary gap discontinuity map. In our analysis, we have empirically chosen a threshold $\mu_{GD}(x, y) > T_d$ where $T_d = 0.5$. Fig. 2 (b) shows an illustration of a gap discontinuity map.

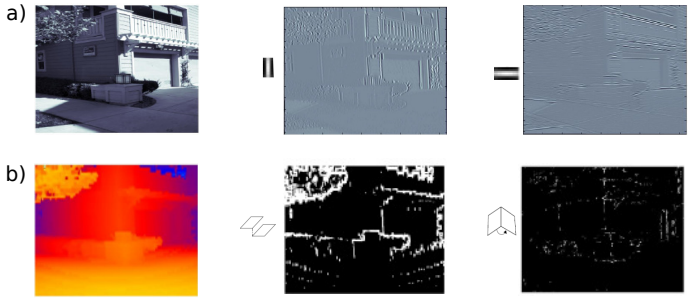

Fig. 2. Examples for features in luminance and depth images. **a)** A gray-scale image convolved with two Gabor filters selective for the same spatial frequency, but different orientation. **b)** A depth map (left) decomposed into its discontinuity maps: gap discontinuity map (middle) and orientation discontinuity map (right).

Surface Orientation Discontinuity. An orientation discontinuity is present when two surfaces meet with significantly different 3D orientations. Orientation discontinuity was measured using surface normal analysis. Here, we considered the methods presented in [1,19]. The orientation discontinuity measure μ_{OD} is computed as the maximum angular difference between adjacent unit surfaces normal. First, a three dimensional point cloud was constructed from the X, Y, Z coordinates for each pixel in a depth image. Then, each pixel is represented by a pixel patch $P_{(x,y,z)}$ compiled from the eight neighboring points in the point cloud. Finally, the unit surfaces normal are computed for each patch $P_{(x,y,z)}$ using Singular Value Decomposition (SVD).

More specifically, for an image patch $P_{(x,y,z)}$ the orientation discontinuity is defined as

$$\mu_{OD}\left(P_{(x,y,z)}\right) = \max\left\{\alpha\left(normal\left(P_{(x,y,z)}\right), normal\left(P_{(x+i,\, y+j,\, \text{z+k})}\right)\right)\right\} \tag{2}$$

where $-1 \leq i, j, k \leq 1$ and $normal\left(P_{(x,y,z)}\right)$: is a function, which computes the unit surface normal of a patch $P_{(x,y,z)}$ in 3D coordinates using Singular Value Decomposition (SVD), α is a function computing the angle between adjacent unit surfaces normal. It is given by

$$\alpha\left(P_1, P_2\right) = \arccos\left(normal\left(P_1\right) \cdot normal\left(P_2\right)\right). \tag{3}$$

max is function to compute the maximum angular difference between adjacent unit surfaces normal. This measure is also thresholded, but based on two criteria, namely i) an *angular criterion*: the maximum angular difference between adjacent unit surfaces normals should be more than a threshold $T_{\theta 1}$ and less than $T_{\theta 2}$, and ii) a *distance-based criterion*: the maximum difference in depth between a point and its eight neighbor's μ_{GD} should be less than a threshold T_d.

In our analysis, we have empirically chosen $T_{\theta 1} = 20°$, $T_{\theta 2} = 160°$ and $T_d = 0.5$, respectively. Fig. 2b shows an illustration of an orientation discontinuity map.

2.5 Classifiers for Predicting Gaze Locations

Opposed to previous computational models that combine a set of biologically plausible filters together to estimate saliency maps, we use a learning approach to train a classifier directly from human eye tracking data. We use a linear Support Vector Machine (SVM) to find out which features are informative. We used models with linear kernels because it performed well for our specific task. Linear models are also faster to compute and the resulting weights of features are easier to understand. We divided our set of images into training images and testing images in order to train and test our model. From each image we chose 200 positively labeled pixels randomly from the top 40% salient locations of the human ground truth saliency map and 200 negatively labeled pixels from the bottom 60% salient locations. In order to have zero mean and unit variance we normalized the features of our training set and used the same normalization parameters to normalize our test data.

For each image in our dataset, we predict the saliency per pixel using a particular trained model. We used the value of $w^T x + b$ (where w and b are learned parameters and x refers to the feature vector) as a continuous saliency map which indicates how salient each pixel is. Then we threshold this saliency map at 40% percent of the image for binary saliency maps.

2.6 Error Measure

The Kullback–Leibler (KL) divergence was used to measure the distance between distributions of saliency values at human vs. random eye positions. We used KL because KL is sensitive to any difference between the histograms, where other measures essentially calculate the rightward shift of histogram1 relative to the histogram2. Also KL is invariant to reparameterizations, such that applying any continuous monotonic nonlinearity to estimated saliency map values[2]. Let $ti = 1 \cdots N$ be N human eye positions in the experimental session. For a saliency model, Estimated Saliency Map is sampled at the human saccade $X_{i,Human}$ and at a random point $X_{i,random}$. First the saliency magnitude at the sampled locations is normalized to the range [0,1]. Then histogram of these values in q=10 bins across all eye positions is calculated. $\Pr(X_{Human}(i))$ and $\Pr(X_{random}(i))$ are the fraction of points in bin i for salient and random points. Finally the difference between these histograms was measured using KL divergence is:

$$KL(X_{Human}; X_{random}) = \sum_{i}^{q} \Pr(X_{Human}(i)) \log\left(\frac{\Pr(X_{Human}(i))}{\Pr(X_{random}(i))}\right). \quad (4)$$

Models that can better predict human fixations show higher KL divergence.

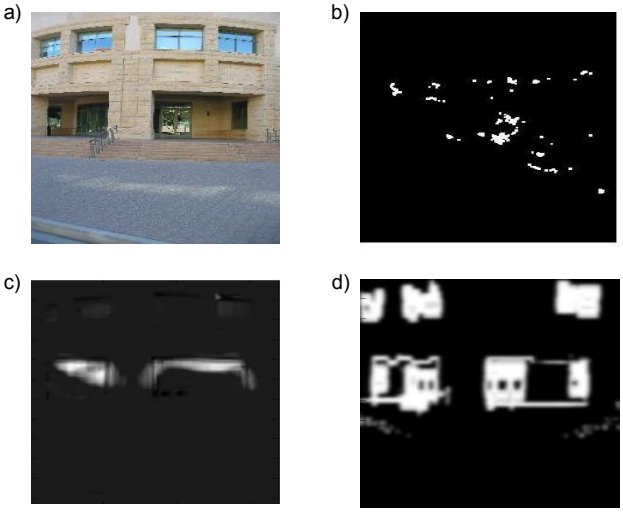

Fig. 3. Examples for features in luminance and depth images. **a)** Natural scene. **b)** Fixation map recorded with our stationary setup . **c)** Itti & Koch features. **d)** Depth discontinuity features.

3 Results 1: Depth Features at the Center of Gaze

We recorded eye movements data from subjects as they viewed static images presented on a computer monitor (see section 2). For each depth image we extracted square image patches around the subject's center of gaze. We also extracted image patches selected at random positions.

3.1 Depth Values around Gaze

We first compared the distribution of depth values of patches in the center of gaze to that expected from random sampling. It is clear that, the distribution of depth values of patches at the center of gaze statistically differ than from random sampling. Figure 4 (a) shows that the normalized histogram of the random sampling from 40 scenes, averaged over all subjects, differ than the distribution of patches in the center of gaze (see Figure 4 (b)) (with P-value = 1.091e-016 of the two-side Kolmogorov–Smirnov (K-S) test with significance level of 0.05).

Figure 4(c) shows that the normalized histogram of patches in the center of gaze over 40 scenes averaged over all subjects in the first three seconds of viewing the scenes differ than the last seven seconds (see Figure 4(d)) (with P-value = 8.6504e-065 of the two-side Kolmogorov–Smirnov (K-S) test with significance level of 0.05). We repeated the statistical test with a maximum of 50m depth and the results was validated.

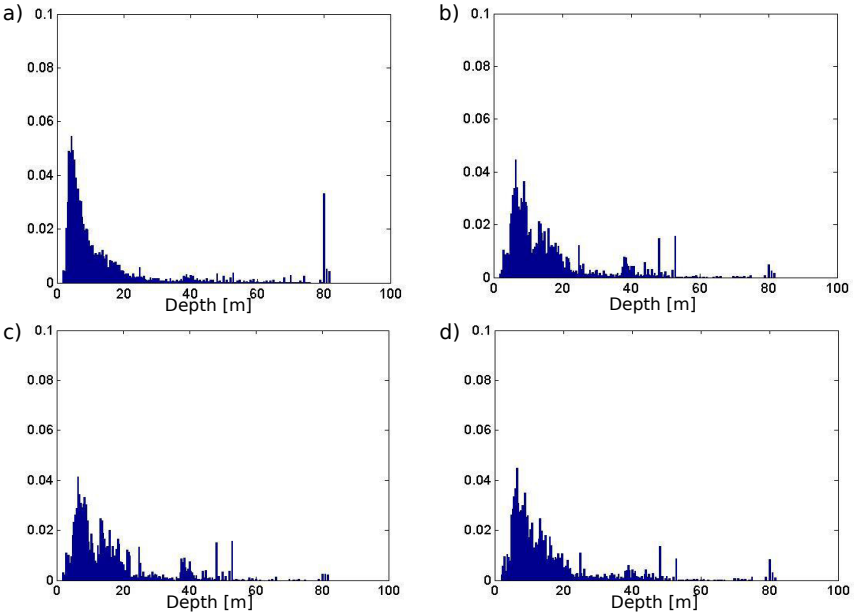

Fig. 4. a) Normalized histogram of depth values of random sampling over 40 scenes, averaged over all subjects. **b)** Normalized histogram of depth at gaze locations, averaged over all subjects. **c)** Normalized histogram of patches in the center of gaze over 40 scene for each subject in the first three seconds of viewing the scenes, averaged over all subjects. **d)** Normalized histogram of patches in the center of gaze over 40 scenes in the last seven seconds of viewing the scenes, averaged over all subjects.

3.2 Depth Features around Gaze

Before we used depth features as new information for predicting eye movements. We examined the presence of depth features around gaze locations. The result of the distribution of depth features in a different neighborhoods around the gaze location averaged over all subjects are shown in Figure 5(a) and the distribution of depth features around gaze for individual subjects are shown in Figure 5(b). It is clear that the presence of depth features around gaze locations are high. This suggest that saliency maps models respecting this will ultimately outperform saliency maps computed only on the basis of 2D pixel images in terms of predicting eye movements.

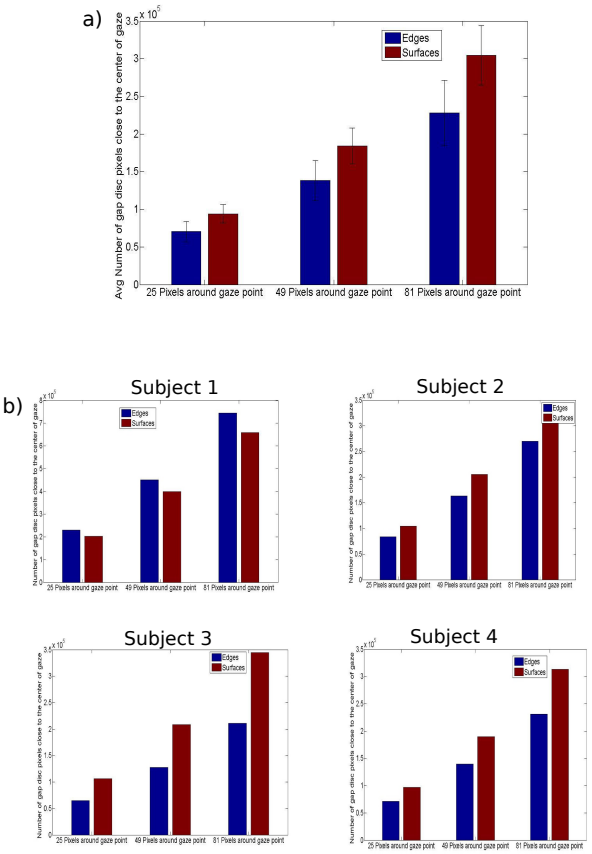

Fig. 5. The presence of depth features in a different neighborhoods around the gaze points. **a)** Bar plot for the presence of depth features in a different neighborhoods around the gaze points, averaged over all subjects. **b)** Bar plot for for the presence of depth features in a different neighborhoods around the gaze points for individual subjects.

4 Results 2: Gaze Location Prediction When Viewing Photos of Natural Scenes

We measured the performance of saliency models using KL divergence (see Section 2.6). Figure 6 describing the performance of different features models for each subject averaged over all testing images. For each image we predict the saliency per pixel using a specific trained model. We can see that the prediction differ according to the type of features we selected. While the model trained on competing saliency features from Itti and Koch perform better than the models trained on other individual features (i.e. only Gabor or only depth features). The averaged result over all subjects shows this finding (see the diagonal of Figure 7).

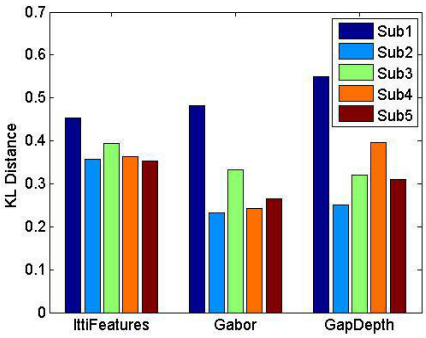

Fig. 6. The KL divergence describing the performance of different SVMs trained on each feature individually, for individual subject

Interestingly the models trained on Itti & Koch combined with depth features outperform models trained on other individual features (i.e. only Gabor or only depth features), or trained on combination of these features. (see Figure 7). It is interesting to note that, depth features combined with luminance features improve the prediction of gaze locations.

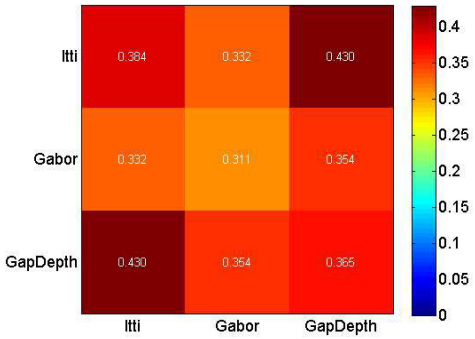

Fig. 7. The KL divergence matrix describing the performance of different SVMs models trained on set of features individually and pairs of features combined, averaged over all subjects. The main diagonal shows the performance of the models trained on individual features. The lower/ upper triangular parts of the matrix show the performance of the models trained on pairs of features combined.

Finally, the overall summary of our analysis is shown in Figure 7 where we computed the KL performance for SVMs trained with different individual features and combined together, averaged over all subjects. We perform the statistical test (t-test2) for all pairs of features (i.e. KL_Itti vs KL_Gabor, KL_Itti

vs KL_GapDepth and KL_Gabor vs KL_GapDepth) with significance level of 0.05 the corresponding P-values were (0.3740, 0.9240 and 0.4488) respectively.

In Figure 7, we see the KL divergence matrix describing the performance of different SVMs models averaged over all subjects. The KL divergence matrix are symmetric with respect to the main diagonal. The main diagonal shows the performance for SVMs models trained on individual features. The lower/ upper triangular parts of the matrix show the performance for SVMs models trained on pairs of features combined.

5 Conclusion

We have analyzed the statistical of depth features in natural natural scenes at the center of gaze. We found that the distribution of depth values of patches at the center of gaze differ than from random sampling. Most interestingly, we found that the presence of depth features around gaze locations were high. This finding points us towards including depth cues into the computation of saliency maps as a promising approach to improve their plausibility.

We also used machine learning to train a bottom-up, top-down model of saliency based on 2D and depth features. We found that models trained on Itti & Koch and depth features combined outperform models trained on other individual features (i.e. only Gabor filter responses or only depth features), or trained on combination of these features. As a consequence, depth features combined with Itti & Koch features improve the prediction of gaze locations.

Our approach, of using joint luminance and depth features is an important step towards developing models of eye movements, which operate well under natural conditions such as those encountered in HCI settings.

Acknowledgments. This work was supported by the DFG GRK 1424 MuSAMA. We also acknowledge the dataset of Laser depth data and luminance Images taken from the Make3D project [15,16].

References

1. Hoover, A., Jean-Baptiste, G., Jiang, X.: An experimental comparison of range image segmentation algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 18, 673–689 (1996)
2. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. 35(1), 185–207 (2013)
3. Gao, D., Vasconcelos, N.: Discriminant saliency for visual recognition from cluttered scenes. In: NIPS (2004)
4. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems 19, pp. 545–552. MIT Press (2007)
5. Horvitz, E., Kadie, C., Paek, T., Hovel, D.: Models of attention in computing and communication: From principles to applications (2003)

6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
7. Lewis, J.P.: Fast normalized cross-correlation (1995)
8. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. IEEE Trans. Pattern Anal. Mach. Intell. 32(1), 171–177 (2010)
9. Mohammed, R.A.A., Schwabe, L.: Scene-dependence of saliency maps of natural luminance and depth images. In: Fifth Baltic Conference "Human - Computer Interaction" (2011) (to appear)
10. Mohammed, R.A.A., Schwabe, L.: A brain informatics approach to explain the oblique effect via depth statistics. In: Zanzotto, F.M., Tsumoto, S., Taatgen, N., Yao, Y. (eds.) BI 2012. LNCS, vol. 7670, pp. 97–106. Springer, Heidelberg (2012)
11. Mohammed, R.A.A., Mohammed, S.A., Schwabe, L.: Batgaze: A new tool to measure depth features at the center of gaze during free viewing. In: Zanzotto, F.M., Tsumoto, S., Taatgen, N., Yao, Y. (eds.) BI 2012. LNCS, vol. 7670, pp. 85–96. Springer, Heidelberg (2012)
12. Potetz, B., Lee, T.S.: Statistical correlations between 2d images and 3d structures in natural scenes. Journal of Optical Society of America, A 7(20), 1292–1303 (2003)
13. Reinagel, P., Zador, A.M.: Natural scene statistics at the centre of gaze. Network 10(4), 341–350 (1999)
14. Roda, C.: Human Attention in Digital Environments. Cambridge University Press, Cambridge (2011)
15. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: NIPS 18. MIT Press (2005)
16. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. 31(5), 824–840 (2009)
17. Durand, F., Judd, T., Ehinger, K., Torralba, A.: Learning to predict where humans look. In: ICCV (2009)
18. Yang, Z., Purves, D.: Image source statistics of surfaces in natural scenes. Network: Computation in Neural Systems 14(3), 371–390 (2003)
19. Yokoya, N., Levine, M.D.: Range image segmentation based on differential geometry: A hybrid approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(6), 643–649 (1989)