PREDICTING SALIENCY MAPS FOR ASD PEOPLE

Alexis Nebout, Weijie Wei, Zhi Liu, Lijin Huang, Olivier Le Meur

PREDICTING SALIENCY MAPS FOR ASD PEOPLE

Alexis Nebout[(1)], *Weijie Wei*[(2,3)], *Zhi Liu*[(2,3)], *Lijin Huang*[(2,3)], *and Olivier Le Meur*[(1)]

[(1)]University of Rennes, CNRS IRISA
[(2)]Shanghai Institute for Advanced Communication and Data Science, Shanghai University, China
[(3)]School of Communication and Information Engineering, Shanghai University, China
alexis.nebout@irisa.fr, codename1995@shu.edu.cn, liuzhisjtu@163.com, hyrx@live.cn, olemeur@irisa.fr

ABSTRACT

This paper presents a novel saliency prediction model for children with autism spectrum disorder (ASD). We design a new convolution neural network and train it with a new ASD dataset. Among the contributions, we can cite the coarse-to-fine architecture as well as the loss function which embeds a regularization term. We also discuss about some data augmentation methods for ASD dataset. Experimental results show that the proposed model performs better than 6 models, one supervised model finetuned with the ASD dataset. Contrary to control people, our results hint that no center bias apply in visuall attention for autistic children.

Index Terms— saliency map, ASD, prediction, visual attention

1. INTRODUCTION

People with autism spectrum disorder (ASD) apprehend differently natural scenes in comparison to neurotypical people [1]. As the ASD diagnosis may require a long and difficult subjective procedure, relying on behavioural, historical and parent-report information [2], new bio-markers are required to identify potential ASD patients. In the recent years, and with the emergence of low-cost eye tracking devices, more and more studies investigate the peculiarities of ocular movement of control vs ASD people. Eye tracking techniques turn out to be fundamental to monitor the gaze deployment in a non-invasive fashion and to reveal deficits related to social cognition such as facial recognition, difficulty of eye contact, to name a few [3, 4, 5].

The recent advent of deep learning techniques show tremendous progress in the ability to predict where an observer stares within a scene has made huge progress. New deep-learning-based saliency models [6, 7] perform much better than non-supervised models such as [8, 9, 10]. However, all these methods have been designed for predicting the salience induced by natural scenes and for healthy people. There has been very few attempts to specialize saliency models for specific cases. We can however cite the saliency prediction for children [11, 12]. In the context of ASD, some recent studies proposed deep networks for saliency prediction [13, 14]. The proposed paper is in the continuity of these studies. We aim to leverage a new ASD datasets by training a new deep network.

The paper is organized as follows. Section 2 presents the architecture of the proposed saliency network. Section 3 elaborates on the training strategy and presents the performances for the proposed method. Section 4 concludes the paper.

This work was supported by the National Natural Science Foundation of China under Grant No. 61771301.

2. PROPOSED DEEP NETWORK

2.1. Architecture

The proposed architecture, inspired from 3 previous saliency models, namely CASNet model [15], deep gaze network [16] and the multi-level deep network of [17], is based on a two-stream VGG-16 network architecture [18]. Figure 1 presents the architecture of the proposed model. The model takes as input RGB images. The first stream extracts high-resolution deep features of images (400×300 pixels) which represent fine-scale information. The second stream is used for coarser-scale images 200×150 for extracting low-resolution deep features, which account for contextual information. For both streams, we extract 1280 feature maps from layers conv3_pool, conv4_pool, conv5_3. Feature maps of layers conv4_pool and conv5_3 are rescaled to get feature maps with a similar spatial resolution. For each stream, the 1280 feature maps go through a 2D convolutional layer (kernel=3×3, stride=1, output=128 maps), a pyramid of dilated convolution (kernel=3×3, stride=1, dilation factors=$\{1, 3, 6, 12\}$, output=4×32) and a 2D convolutional layer (kernel=3×3, stride=1, output=32 maps). The fine and coarse-based maps (2×32) are then concatenated. Following the proposition of [15], a 2×2 max pooling is applied on the 64 channels of concatenated feature maps to reduce their spatial variance. We then use a 2D locally convolutional layer to determine the relative importance of the different maps. These maps are then used to weight the 64 channels of concatenated feature maps by a pixel-wise multiplication. A saliency map is finally determined thanks to a 1×1 convolutional layer. The different layers have a Relu activation.

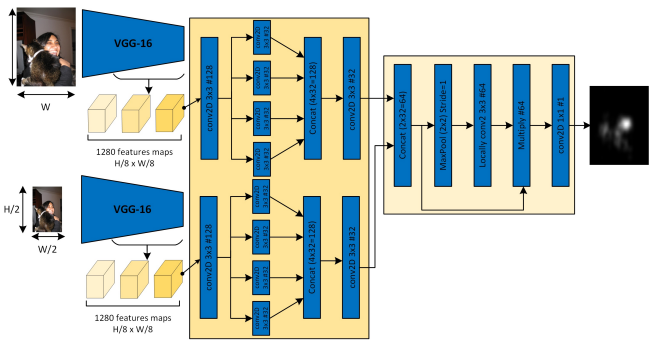

Fig. 1: Proposed deep architecture.

2.2. Optimization and loss function

The network was trained using stochastic gradient descent. To prevent over-fitting, a dropout layer is added in each stream before the first 2D convolutional layer. The rate of dropout is set to 0.25. During training, the network was validated against the validation set after every iterations to monitor convergence and over-fitting. The learning rate is set to 0.001. To test the model, the predicted map is first resized to get the original image resolution and is then filtered by a Gaussian filter to smooth the generated saliency map.

The loss function $\mathcal{L}$ between the ground truth map S and the prediction $\hat{S}$ is similar to the loss used in [17]:

$$\mathcal{L}(S,\hat{S}) = \frac{1}{N \times M} \sum_{i=1}^{N \times M} \left(\frac{1}{\alpha - S_i} (S_i - \hat{S}_i)^2 + \beta \times R_i \right) \quad (1)$$

with, $alpha = 1.1$ and R_i a regularization term as shown in the equation 2 below. We assume that S and $\hat{S}$ are in $[0, 1]$. $N \times M$ represents the image resolution. We empirically set the parameter β to 0.1, even though it could be optimized to improve final results.

$$R_i = (\hat{S}_i - B_i)^2 \quad (2)$$

The map B, as illustrated in Figure 2 (right), represents the positional bias observed with ASD people. The map B is simply the average saliency maps computed over the training dataset. Figure 2 also illustrates on the left-hand side the positional bias for people without ASD. We observe a strong difference between both populations. The center bias is much less important for people with ASD than for the control population.

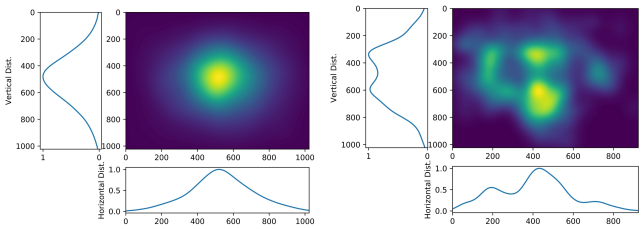

Fig. 2: Averaged colored saliency maps computed over the training dataset for people without ASD (left) and with ASD (right). Horizontal and vertical marginal distributions are also plotted.

The total trainable parameters of the proposed model is 65 357 441. Compared to existing models, this number is quite reasonable. For instance, the very recent CASNet model [15] requires more than 142 million of trainable parameters whereas Sam-VGG and Sam-ResNet [6] require between 50 and 70 million of trainable parameters.

3. EXPERIMENTS

3.1. Training, validation and test datase

ASD dataset is presented in [19]. It is composed of 300 images selected among the MIT1003 dataset. The experiment was carried on 14 autistic children and 14 typically developping (TD) children. ASD children were between 5 and 12 years old with a mean of 8 years old. Both TD and ASD population were 8 years old in average.

We used the first 240 images of the ASD dataset for the training and fine-tuning the proposed method. The next 30 images are used for validation whereas the last 30 images are used for testing.

Fig. 3: Different data augmentation methods. From left to right: original, blurred, flipped, noisy and grayscale image.

3.2. Data augmentation

In order to get more training samples, we augment the data in different ways as illustrated in Figure 3. We assume that used transformations do not significantly modify gaze behaviors:

- Horizontal flip: this transformation allows to flip an image right/left (e.g. mirror image).
- Blur: we add some blur to images by considering that it does not change the visual gaze deployment [20].
- Noise: we add some noise to images assuming that it does not change the visual gaze deployment.
- Grayscale: we convert color images into grayscale images.

Based on the proposed architecture illustrated in Figure 1, we define and assess different variants of the proposed method. These variants are described below. They are composed of two main parts: the first one is related to a training process from a random initialization whereas the second concerns a fine-tuning process.

- **Random init. (RI)**: We randomly initialize the weights of the network. Then, we augment the training dataset according to 3 methods. **Method M1** implements the horizontal flip as data augmentation, while **method M2** also adds blurred images. **Method M3** combines flip, blur, noise and grayscale data augmentations.
- **Fine-tuning (FT)**: We pre-trained the network with Judd dataset. The model is then fine-tuned with the ASD dataset. We use the loss $\mathcal{L}$ plus the regularization term (see equation 1 and 2). The first **method M1** implements the horizontal flip as data augmentation, while **method M2** adds blurred images. We also modify the loss $\mathcal{L}$ of **M2** in order to leverage the complementary between MLNET loss, Kullback-Leibler loss and the regularization term.

3.3. Performances

To carry out the evaluation, we use quality metrics used in the MIT benchmark [21]: CC (correlation coefficient, $CC \in [-1, 1]$), SIM (similarity, intersection between histograms of saliency, $SIM \in [0, 1]$), AUC (Area Under Curve, $AUC \in [0, 1]$), NSS (Normalized Scanpath Saliency, $NSS \in]-\infty, +\infty[$) and KL (Kullback Leibler divergence, $KL \in [0, +\infty[$). Details of these metrics can be found in [21, 22].

Figure 4 illustrates an example of saliency maps predicted by the different models. Table 1 presents the results of the proposed methods as well as the performance of existing methods. Several observations can be made.

How do existing saliency models perform? Seven existing models are first evaluated. As expected, non-supervised models, such as Rare2012 [8], Hou [23], AWS [10], GBVS [24], SUN [9] do not perform well. The SAM-VGG [6] performs better than unsupervised model, but its performances are lower than the proposed methods, as we shall see. The SalGAN model [7], when fine-tunes with the training ASD dataset, presents the best performances among the existing models.

Performance of the proposed methods. The best results are obtained by the method FT-M1. The best 2 and worse 2 predicted

Fig. 4: Predicted saliency maps for the different tested saliency models. From left to right, first row: original image and ground truth; second row presents the predictions from RARE2012, Hou, AWS, SUN, GBVS, SAM-VGG. From left to right, third row presents the predictions from SalGAN, RI-M1, RI-M2, RI-M3, FT-M1, FT-M2. The red frame indicates the prediction having the highest CC.

Table 1: Performance of the proposed model and comparison with existing saliency models. SalGAN$^{(*)}$=SalGAN model finetuned by MIT1003 and training dataset. RI= Random Initialization; FT=Fine-Tuning. The last line present the results of FT-M2 on test set of 200 natural images from [19]. Results on bold show best scores, while results on italic in the last line show the greater score on the other dataset.

Model	SIM ↑	CC ↑	KL↓	NSS ↑	AUC-J ↑	AUC-B ↑
Existing models						
RARE2012	0.5317	0.4240	0.7754	0.8632	0.7224	0.7058
Hou	0.5304	0.3934	0.7127	0.7452	0.7088	0.6940
AWS	0.5178	0.3777	0.8024	0.7551	0.6973	0.6897
SUN	0.4834	0.2442	0.8842	0.5144	0.6376	0.6301
GBVS	0.5990	0.5541	**0.5426**	0.9919	0.7642	0.7555
SAM-VGG	0.5453	0.5961	3.3719	**1.3182**	0.7758	0.6576
SalGAN$^{(*)}$	0.6353	0.6866	1.5651	1.3074	0.7829	0.7551
Proposed methods						
RI - M1	0.6237	0.6808	2.5995	1.2709	0.7833	0.7520
RI - M2	0.6211	0.6587	2.0173	1.2140	0.7810	0.7540
RI - M3	0.5833	0.5655	1.9500	0.9714	0.7573	0.7297
FT - M1	**0.6590**	**0.6983**	0.9480	1.2637	**0.7955**	**0.7739**
FT - M2	0.6099	0.5883	0.6368	1.0274	0.7712	0.7468
Results on other dataset						
FT - M1	0.6308	0.6822	0.9023	*1.4193*	*0.8106*	*0.7850*

saliency maps are illustrated on figure 5. Among the different setups, we observe that the blur augmentation method decreases SIM, CC, KL, NSS and AUC-J. It may be correlated with the fact that ASD people give more importance to details in natural scenes [25]. The third augmentation supports this claim, especially with noise. Furthermore, our results on RI are worse than SalGAN, or at least similar. On the other hand, the increase in SIM, CC, AUC-J and AUC-B scores on the two last lines indicates that the fine-tuning of the proposed method improves the performances; in this context, the model outperforms the state-of-the-art models. The blur augmentation employed in RI-M2 may be the cause of the drop in performance, supporting the point stated above.

Comparing FT-M1 method with fine-tuned SalGAN. In Table 2, we further analyze the performance of the best proposed model per image of the test dataset. For each image, we compare the correlation coefficient of the best proposed method and the SalGan (fine-tuned on ASD dataset) model. Some annotations are also given indicating what are the most important elements in the scene.

Do predicted saliency maps present the same positional bias as ground truth ones? We also assess whether or not we retrieve a positional bias, as the one illustrated in Figure 2 (right). For that, we average the 30 predicted saliency maps. Figure 6 illustrates the

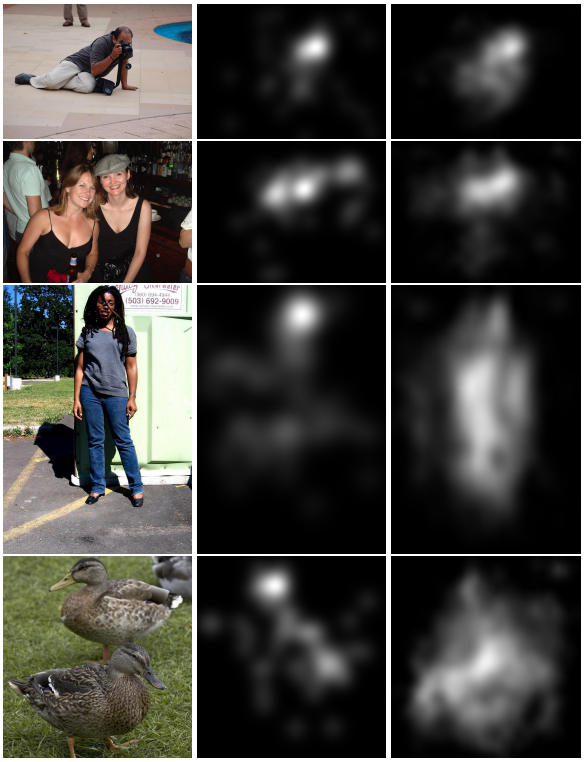

Fig. 5: Predicted saliency maps computed by the best performing method FT-M1 method. From left to right: original image, ground truth saliency map, and predicted saliency map. From top to bottom: top 2 predictions with a CC score of 0.8992 and 0.8653, respectively. The last two rows present the worst prediction with a CC score of 0.4869 and 0.4295, respectively.

bias of our predicted saliency maps. We observe that the positional bias of predicted saliency map does not compare well with the one of Figure 2 (right). To make clear this point, we compute the positional bias of the test dataset. It appears that the bias of the test dataset does not present the same bias as the bias of the training set.

4. CONCLUSION

In this paper, we have presented a novel saliency prediction model dedicated to children with ASD. The proposed coarse-to-fine architecture allows to combine fine details as well as global information. Thanks the training strategy, based on the proposed loss function, the fine-tuning and the data augmentation, the proposed model outperforms existing saliency models. Future works will concern mainly the dataset extension.

5. REFERENCES

[1] A. Klin, "Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism," *Arch Gen Psychiatry*, vol. 59, pp. 809–816, Sep. 2002.

[2] V. Yaneva, L. A. Ha, S. Eraslan, Y. Yesilada, and R. Mitkov, "Detecting autism based on eye-tracking data from web search-

Table 2: CC score on test dataset, according to SalGAN and FT-M1 models. Images are sorted by descending CC on FT-M1.

Image	Image context	CC FT-M1	CC SalGAN	Image	Image context	CC FT-M1	CC SalGAN
295	People	0.8992	0.6005	300	Nature	0.7025	0.7701
292	People	0.8653	0.7963	274	Building	0.7018	0.3562
275	Toilet	0.8235	0.8355	283	People	0.7011	0.8358
291	People	0.8172	0.6836	288	Statue of a man	0.6609	0.6506
296	People	0.7946	0.6650	293	Infrastructure	0.6574	0.6673
281	Street	0.7887	0.7073	278	Bed	0.6506	0.5973
289	People	0.7838	0.6562	277	Bed	0.6179	0.5471
294	People	0.7759	0.7455	290	People	0.6148	0.7256
279	People	0.7629	0.7519	271	Car	0.5981	0.5599
287	People	0.7605	0.8517	299	People	0.5860	0.7881
284	People	0.7603	0.6868	297	People	0.5769	0.5318
280	Street	0.7545	0.5972	272	People	0.5114	0.6697
276	Toilet	0.7529	0.7513	273	People	0.4958	0.5600
298	People	0.7181	0.8246	282	People	0.4869	0.7227
285	People	0.7089	0.8565	286	Ducks	0.4295	0.6067

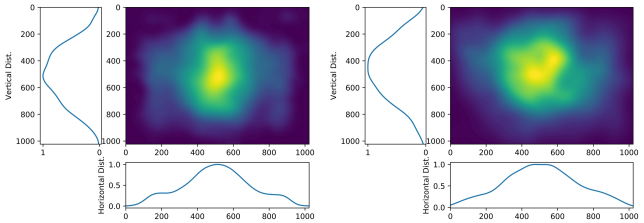

Fig. 6: Positional bias of predicted (left) and ground truth (right) saliency map for ASD people computed over the test dataset.

ing tasks," in *Proceedings of the Internet of Accessible Things*. ACM, 2018, p. 16.

[3] M. T. Mercadante, E. C. Macedo, P. M. Baptista, C. S. Paula, and J. S. Schwartzman, "Saccadic movements using eye-tracking technology in individuals with autism spectrum disorders: pilot study," *Arquivos de neuro-psiquiatria*, vol. 64, no. 3A, pp. 559–562, 2006.

[4] E. Thorup, P. Nyström, G. Gredebäck, S. Bölte, and T. Falck-Ytter, "Altered gaze following during live interaction in infants at risk for autism: an eye tracking study," *Molecular autism*, vol. 7, no. 1, pp. 12, 2016.

[5] G. Dawson, "Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies," *Developmental neuropsychology*, vol. 27, no. 3, pp. 403–424, 2005.

[6] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2018.

[7] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," *arXiv preprint arXiv:1701.01081*, 2017.

[8] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit, "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Processing: Image Communication*, vol. 28, no. 6, pp. 642–658, 2013.

[9] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of vision*, vol. 8, no. 7, pp. 32–32, 2008.

[10] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image and Vision Computing*, vol. 30, no. 1, pp. 51–64, 2012.

[11] O. Le Meur, A. Coutrot, Z. Liu, P. Rämä, A. Le Roch, and A. Helo, "Visual attention saccadic models learn to emulate gaze patterns from childhood to adulthood," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4777–4789, 2017.

[12] O. Krishna, A. Helo, P. Rämä, and K. Aizawa, "Gaze distribution analysis and saliency prediction across age groups," *PloS one*, vol. 13, no. 2, pp. e0193149, 2018.

[13] M. Jiang and Q. Zhao, "Learning visual attention to identify people with autism spectrum disorder," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3267–3276.

[14] H. Duan, G. Zhai, X. Min, Y. Fang, Z. Che, X. Yang, C. Zhi, H. Yang, and N. Liu, "Learning to predict where the children with asd look," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 704–708.

[15] S. Fan, Z. Shen, M. Jiang, B. L Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7521–7531.

[16] M. Kümmerer, T. SA Wallis, and M. Bethge, "Deepgaze ii: Reading fixations from deep features trained on object recognition," *arXiv preprint arXiv:1610.01563*, 2016.

[17] M. Cornia, L. Baraldi, G. Serra, and Rita C., "A Deep Multi-Level Network for Saliency Prediction," in *International Conference on Pattern Recognition (ICPR)*, 2016.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, and P. Le Callet, "A dataset of eye movements for the children with autism spectrum disorder," in *ACM Multimedia Systems Conference (MMSys'19)*, 2019.

[20] T. Judd, F. Durand, and A. Torralba, "Fixations on low-resolution images," *Journal of Vision*, vol. 11, no. 4, pp. 14–14, 2011.

[21] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark," 2015.

[22] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior Research Method*, vol. 45, no. 1, pp. 251–266, 2013.

[23] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[24] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2007, pp. 545–552.

[25] A. Alink and I. Charest, "Individuals with clinically relevant autistic traits tend to have an eye for detail," *bioRxiv*, 2018.