A PRELIMINARY BENCHMARK OF FOUR SALIENCY ALGORITHMS ON COMIC ART

Khimya Khetarpal

University of Florida
Dept. of ECE
Email: kkhetarpal@ufl.edu

Eakta Jain

University of Florida
Dept. of CISE
Email: ejain@cise.ufl.edu

ABSTRACT

Predicting the salient regions of a comic book panel has the potential to drive a variety of applications such as segmentation, cropping, effects such as moves on stills, etc. Computational saliency algorithms have been widely tested on a variety of natural images, and extensively benchmarked. We report the performance of four saliency algorithms on a set of comic panels taken from public domain legacy comics. We find that a data-driven method performs highest based on two metrics, Normalized Scanpath Saliency and Area Under the Curve. We discuss possible reasons for this finding based on an exploratory analysis of the similarity between the comic images in our dataset and images used in the dataset of the data driven method.

***Index Terms*—** saliency, human gaze, predicting fixations, comic art

1. INTRODUCTION

Knowing the salient components in an image is useful for a variety of applications, such as compression, segmentation, auto cropping, etc. Comic artists and filmmakers use visual symbols as cues to lead the viewer's attention [1]. Understanding saliency in comic art could illustrate design decisions such as effective color contrast, the placement of objects and word bubbles, captions, illumination, etc. The two widely accepted approaches to model visual saliency are bottom-up and top-down approaches. The former is a stimulus driven approach, according to which features in an image that are very distinguishable by the human eye stand out [2]. Top-down models take a task-driven approach, driven primarily by the goal of viewing the image, for example, finding all occurrences of a flower in an image, or, trying to memorize different terms of a human heart [3].

Our contributions We perform a comparative analysis of four saliency models on a set of comic panels taken from public domain legacy comics. These saliency models are a data-driven approach (LSVM) [4], a graph based bottom-up saliency model (GBVS) [5], a difference of gaussian based bottom up algorithm (VOCUS2) [6] and a region contrast based salient object detection (RC) [7]. Our primary contribution is benchmarking these saliency models for comic art. We find that LSVM outperforms other three models on comic images. We discuss possible reasons for this, by exploring the similarity between comic art and natural images used by LSVM for training and testing.

2. RELATED WORK

Saliency modeling has been an active field of research. The earliest models used selective features at several image scales to form a bottom-up saliency map ([8], [9] and [10]). [5], [11], [12], [13], [14] are several approaches based on the bottom-up technique of modeling saliency. Top-down saliency is driven by tasks, rewards, emotions and expectations, as opposed to physical characteristics of bottom-up detection [15], [16], [17]. Recent advances in predicting fixations use deep neural networks [18], [19], [20], [21].

Based on consistent success over time, LSVM and GBVS form two natural choices for this benchmark study. VOCUS2 and RC have been chosen because they are recent models that report outperforming several other models in different applications. LSVM, a data-driven approach, combines low, mid and high level features, and trains a Linear Support Vector Machine on gaze data from multiple viewers, performing a free viewing task on several images. GBVS, a non data-driven approach, aims at exposing connected regions of dissimilarity, defined as the distance between the intensity of two pixels, measured on a logarithmic scale. VOCUS2, a derivative of the original, biologically inspired saliency model in [8] employs a pyramid structure consisting of twin pyramids with multiple scales per layer as opposed to one pyramid with one scale per layer used originally. RC [7] uses a graph-based image segmentation technique to form regions in an input image, and computes region level saliency by finding color contrast to all other regions in that image.

3. EYE TRACKING DATA ON COMIC IMAGES

Our comic dataset consists of 23 images, taken from public domain legacy comics. We recorded eyetracking data from

978-1-4799-7082-7/15/$31.00©2015 IEEE

5 viewers (3 male), using a remote eyetracker (SensoMotoric Inc, SMI RED-m, 120 Hz). Participants were seated at a distance of approximately 64 cm from the screen (1680 × 1050) resolution, 18in×11in. A visual angle of 1 degree is approximately 37 pixels at these settings. Data consisted of fixations in the x and y space coordinates, event information comprising either of a 'blink', 'fixation', or, a 'saccade', along with the stimulus name, with corresponding timestamps. Figure 1 depicts a few samples from the comic dataset, where the gaze data of 5 subjects has been overlaid on the stimulus. Each subject's fixations are marked in a different color. Samples where the gaze position was outside the screen were discarded.

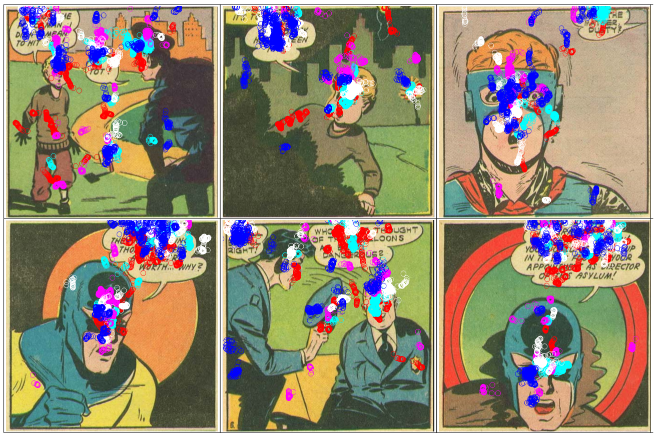

Fig. 1: Gaze data from 5 participants shown as different colored circles on randomly chosen images from comic panels.

4. BASELINE CHECK

To validate our test setup, we perform a comparative analysis for all four models over three image categories from the CAT2000 dataset [22]: Outdoor Man-made, Outdoor Natural, and Social. We perform experiments to evaluate the ability of models to predict ground truth human fixations. [23] reports the evaluation of submitted models with several metrics, such as, Similarity Measure (SIM), Earth Mover's Distance (EMD), Pearson's Linear Coefficient (CC), Normalized Scanpath Saliency (NSS), and different versions of Area Under Curve (AUC). Their results are based on their evaluation on all 20 categories and ground truth from 24 observers. We compare the performance of LSVM and GBVS on our 3 category test setup and compare the scores against those reported by [23] as a sanity check on our test setup.

Table 1: Mean NSS Scores reported by our test setup on three categories from natural images show similar trend when compared to the scores reported by [23] for the same image categories. Standard deviations are reported in brackets.

Our Test Setup				
Normalized Scanpath Saliency [NSS]	**Outdoor ManMade**	**Outdoor Natural**	**Social**	**Overall**
LSVM	1.22 (0.22)	1.28 (0.21)	1.21 (0.19)	1.24 (0.21)
GBVS	1.11 (0.28)	1.17 (0.34)	1.14 (0.27)	1.14 (0.3)

Benchmark Reported Scores				
Normalized Scanpath Saliency [NSS]	**Outdoor ManMade**	**Outdoor Natural**	**Social**	**Overall**
LSVM	1.27 (0.2)	1.23 (0.21)	1.18 (0.19)	1.23 (0.2)
GBVS	1.18 (0.34)	1.06 (0.36)	1.1 (0.32)	1.11 (0.34)

Table 2: Our test setup and the benchmark [23] report similar AUC scores for natural image categories namely, Outdoor Natural, Outdoor ManMade and Social.

Our Test Setup				
Area Under Curve[AUC]	**Outdoor ManMade**	**Outdoor Natural**	**Social**	**Overall**
LSVM	0.83 (0.04)	0.84 (0.04)	0.83 (0.05)	0.83 (0.04)
GBVS	0.79 (0.06)	0.79 (0.05)	0.79 (0.05)	0.79 (0.05)

Benchmark Reported Scores				
Area Under Curve[AUC]	**Outdoor ManMade**	**Outdoor Natural**	**Social**	**Overall**
LSVM	0.84 (0.04)	0.84 (0.04)	0.83 (0.04)	0.84 (0.04)
GBVS	0.8 (0.05)	0.78 (0.06)	0.79 (0.06)	0.79 (0.06)

Table 1 depicts mean NSS scores from our test setup and those reported by the benchmark. The overall NSS score is averaged across the 3 categories. Similarly, average AUC score for our experiments compared to benchmark values are reported for LSVM and GBVS as shown in the Table 2. Here, it should be noted that the scores from our test setup would not mirror scores from the benchmark [23], because the latter evaluates the models on test images and different observers, whereas our test setup has been evaluated on the publicly available training data. The authors of RC [7] report a comparison of their model with LSVM, and based on the AUC score, LSVM is reported to outperform their model in predicting human fixations. Figures 5 and 6 show LSVM outperforms all other models on all four categories including comic, based on both NSS and AUC scores per our evaluation too.

5. EXPERIMENT

To evaluate the performance of the four saliency algorithms, we use code made available by the authors with default settings. We obtain saliency maps for all 23 comic images. Figure 2 shows saliency maps for six chosen comic panels. Comic panels 2, 6 and 8 are selected based on their high NSS scores for all four models (Figure 4). It is observed that almost all saliency algorithms show relatively low performance for comic panels 7, 14 and, 20. These panels illustrate certain comic images are difficult to detect visual attention with the models in our test setup. Figure 2 depicts regions marked salient by each model on these specific images.

5.1. Performance Analysis

The eyetracking data is overlaid on the comic images resized to the same resolution at which they were displayed at, for data collection from five subjects. Once we obtain the

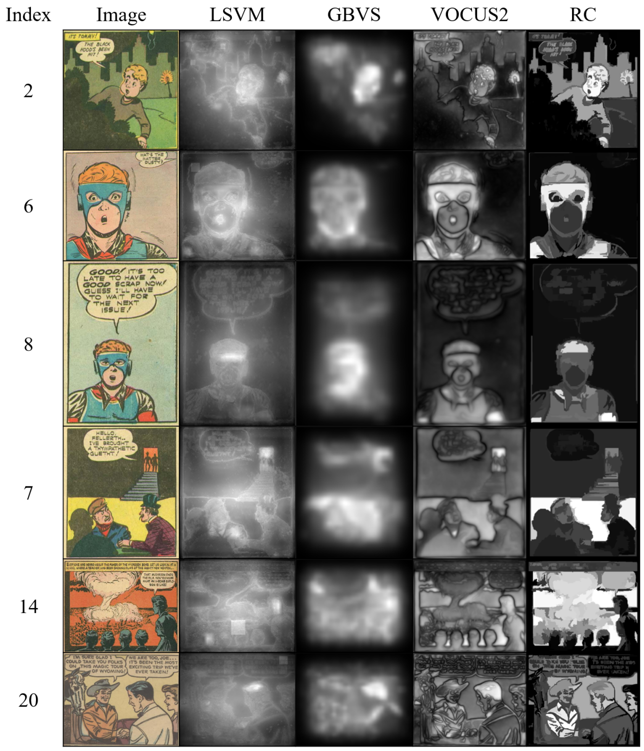

Fig. 2: Saliency maps generated from four different algorithms are depicted for chosen comic images. LSVM, GBVS show center bias, and similar regions are marked salient. VOCUS2 finds it difficult to detect multiple salient regions, whereas, RC marks complete regions as salient.

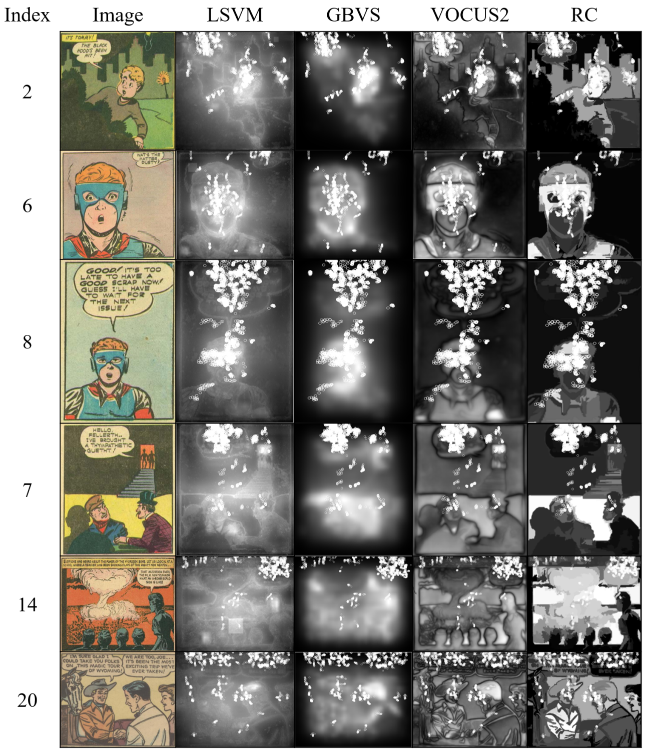

Fig. 3: Gaze data from 5 participants is overlaid using white circles on the saliency maps generated from LSVM, GBVS, VOCUS2 and RC to show the overlap between where humans looked and the regions predicted by the saliency algorithms.

saliency maps from all four models for 23 images from comic dataset, next, we overlay the fixations on the saliency maps as shown in the Figure 3.

To evaluate the performance of each model on comic images, we use Normalized Scanpath Saliency (NSS), as described in [24]. Each saliency map was linearly normalized to have zero mean and unit standard deviation. The normalized saliency values were extracted corresponding to the fixation locations for each subject, and the mean of these values was taken as a measure of the correspondence between the saliency map and scanpath. For each model, mean of all the five subjects' NSS score results in the average NSS score for an image. The same is repeated to compute the mean NSS score for all four models across 23 comic images.

Figure 4 depicts the mean NSS score variation for each model, per image. It is interesting to note similar trends in performance across all models for comic images 2, 6 and 8, as well as 7, 14 and 20. Most models perform well in images 2, 6 and 8, whereas, they find it difficult to detect saliency in images 7, 14 and 20. We observed that GBVS shows peak performance at many instances, which are very close to the values of LSVM. However, LSVM outperforms GBVS, on an average across 23 comic images as shown in black bars and dark gray bars respectively in Figure 5 . We repeat the same experiment for three natural image categories from CAT2000 dataset. Figure 5 depicts an overview of performance of all four models in terms of mean NSS score.

We also report the performance of the models using another metric, Area Under the Curve (AUC). We use the implementation by [25] which requires a binary fixation map. For comic dataset, the raw data has fixations coordinates: x and y for each stimuli. We create a binary map with 1 corresponding to the fixation locations and 0 corresponding to the background locations . The scores reported from these experiments are also shown in Figure 6 . As shown in black bars, LSVM consistently outperforms all other models. GBVS is only marginally below LSVM as observed from dark gray bars. We observe that VOCUS2 performs the lowest, while RC is only marginally better than VOCUS2 for both NSS and AUC scores across all categories.

6. EXPLORATORY ANALYSIS

Our goal is to explore possible reasons for LSVM performing so well on comic images, when it is a data-driven model trained on natural images that (on the first glance) are not "comic like" in style. LSVM was never trained on comic im-

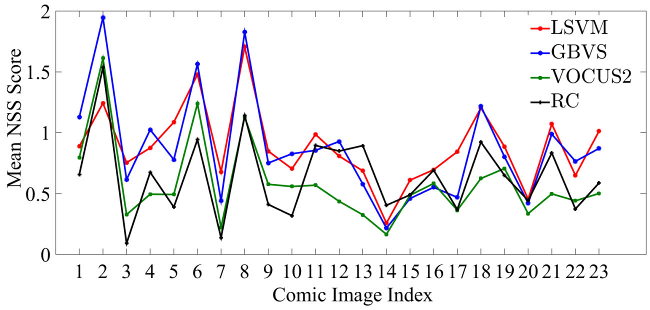

Fig. 4: Mean NSS Scores across 23 images. Comic images 2, 6 and 8 reported a relatively higher NSS scores for all models, whereas, consistent low NSS scores were observed for comic images 7, 14 and 20.

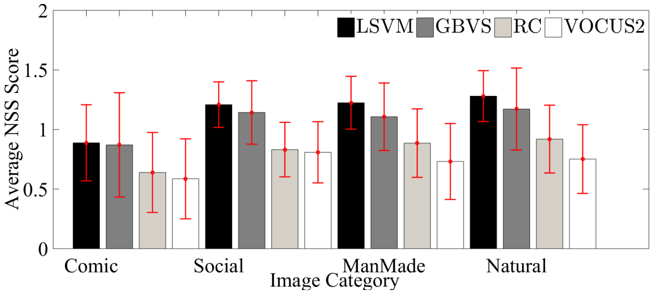

Fig. 5: Mean NSS score where error bars are standard deviation. LSVM consistently outperforms GBVS, RC and VOCUS2. Performance on natural image categories is relatively higher than that on comic images for all models.

ages, yet it outperforms all the other approaches for predicting fixations, not only in context of natural images, but also, for comic art. We present our hypothesis and experiments as supporting evidence to discuss possible reasons.

Hypothesis 1: Comic panels in our dataset are similar to natural images in LSVM's dataset [4] in terms of features used for saliency prediction.

We consider the weights assigned to each of the 33 feature channels in LSVM's model of saliency. We sort these weights. The most weighted features are the coefficients of the second coarsest sub-band of steerable pyramids [26], Viola Jones face detector [27], features used in a saliency model described by Torralba [28] and Rosenholtz [29], and color contrast as calculated by Itti and Koch's saliency method [12], in decreasing order of the weights.

We use the implementation in LSVM to compute feature vectors for 1003 images from the LSVM dataset and 23 images from our comic dataset. We compute the distances between comic and natural images in feature space, using Euclidean distance. Each element in the resulting 23 × 1003 matrix holds the pairwise distance score. These scores are averaged across all comic images, resulting in a 1×1003 vector.

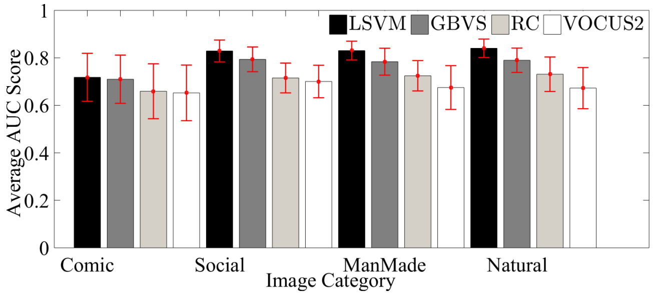

Fig. 6: Mean AUC score with error bars as standard deviation. Across all image categories, LSVM is ranked on top, followed by GBVS, RC and VOCUS2, in decreasing order of accuracy. Of the 4 image categories, all models are reported to show relatively lower performance on comic images.

This represents the average distance between comic and natural images, based on the feature being considered. Similarly, we compute pairwise distances between natural images which are then averaged across all natural images, resulting in a average distance vector of 1×1003 dimension. To quantify similarity between comic images and natural images in a particular feature space, we define a similarity score which can be expressed as,

$$\begin{aligned} Similarity \quad Score = (&Average \quad distance \quad between \\ &natural \quad images) \quad - \\ (&Average \quad distance \quad between \\ &natural \quad and \quad comic \quad images) \end{aligned}$$

With this definition, a high positive score indicates comic images were relatively more similar to natural images than the amount natural images are similar to other natural images. This is observed in context of using the second coarsest sub-band from steerable pyramids as the feature, based on the Figure 7. A low positive score is indicative of the fact that comic images are as similar to natural images as natural images among themselves. Similarly, a negative score depicts high dissimilarity between comic and natural images as compared to natural images. The latter is observed when color contrast as calculated by Itti and Koch's saliency model [12] is used as the feature.

In other words, LSVM is expected to perform relatively the same on comic images as it does on natural images, if the similarity score is a high positive in accordance with our metric. This could explain why LSVM outperforms the other three saliency models that we benchmarked. It turns out that comic images are quite similar to natural images based on the features used by this model. These are preliminary findings because we have only studied 3 of the 33 features and only 23 comic images, but they could point to interesting future work directions. It will be interesting to perform the same experiment on manga images for example.

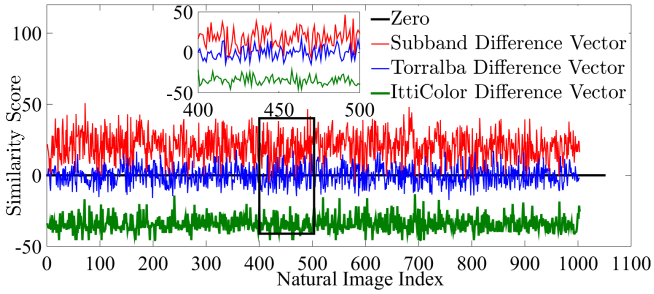

Fig. 7: High positive score using second coarsest sub-band from steerable pyramids as feature, shows higher similarity between comic and natural images, relative to the similarity among natural images. Low positive score indicates similarity between comic and natural images is of the same order as among the natural images, in context of features from Torralba saliency. The plot below the zero line, indicates comic images are lot different from natural images relative to the pool of natural images, in context of color channel as feature.

Hypothesis 2: Face detection algorithms used by saliency models work well for natural images but not for comic art.
Our results and analysis indicate that all four models including LSVM show relatively poor performance for comic art, as compared to the performance of the the same model on natural images (Figures 5 and 6). To understand this finding further, we study features which are highly weighted in LSVM's saliency algorithm, in particular, faces. Faces are not only the second highest weighted feature, but also, one of the most gazed upon objects in an image [4].

We compute the accuracy of face detection for comic images, using the Viola Jones face detector [27]. The ground truth was marked by a human coder. Of the 40 faces in comic images, 7.5% faces were correctly detected, whereas 72.5% faces were false positives. Figure 8 shows a few samples from this experiment. From these results, we observe that the face detection fails for most of the comic images. This indicates that faces could be a potential reason for much lower performance of LSVM on comic art as compared to it's performance on other natural image categories.

6.1. Conclusion

In this paper, we benchmark four saliency algorithms on comic images. We perform a comparative analysis of these algorithms in detecting salient regions in comic panels with eyetracking data. We find that a data-driven approach, LSVM, outperformed three other saliency models included in our test setup. We discuss potential reasons for these findings.

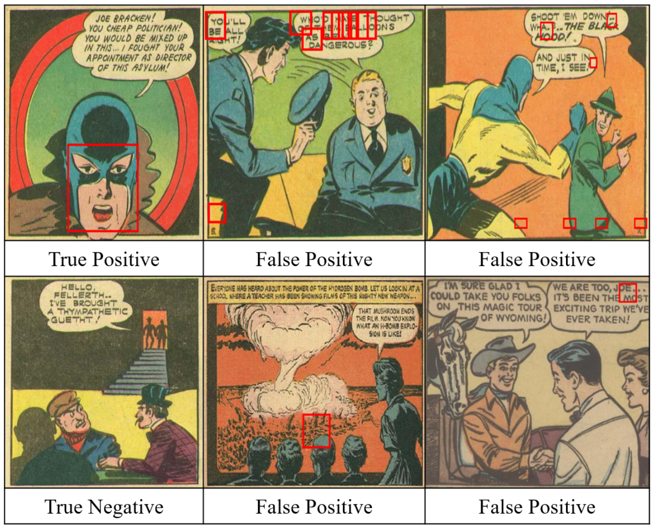
True Positive
False Positive
False Positive
True Negative
False Positive
False Positive
JOE BRACKEN! YOU CHEAP POLITICIAN! YOU WOULD BE MIXED UP IN THIS... I FOUGHT YOUR APPOINTMENT AS DIRECTOR OF THIS ASYLUM!
YOU'LL BE ALL RIGHT!
WHO'D HAVE THOUGHT OF THEM BALLOONS AS BEING DANGEROUS?
SHOOT 'EM DOWN... WHA..?..THE BLACK HOOD!
AND JUST IN TIME, I SEE!
HELLO, FELLERTH.. I'VE BROUGHT A THYMPATHETIC GUETHT!
EVERYONE HAS HEARD ABOUT THE POWER OF THE HYDROGEN BOMB. LET US LOOK IN AT A SCHOOL WHERE A TEACHER HAS BEEN SHOWING FILMS OF THIS MIGHTY NEW WEAPON...
THAT MUSHROOM ENDS THE FILM. NOW YOU KNOW WHAT AN H-BOMB EXPLOSION IS LIKE!
I'M SURE GLAD I COULD TAKE YOU FOLKS ON THIS MAGIC TOUR OF WYOMING!
WE ARE TOO, JOE... IT'S BEEN THE MOST EXCITING TRIP WE'VE EVER TAKEN!

Fig. 8: Of the 40 faces in comic image data set, only 3 faces were detected correctly by Viola Jones face detector, whereas 29 detections were false positives. Faces in comic art are a lot different than those in natural images in style.

Evaluation of deep learning based saliency detection algorithms for predicting fixations could provide a deeper understanding of visual attention in comic art. It would be interesting to see performance of deep learning tools as they are free from hand-crafted features. Improving existing saliency algorithms to detect visual attention in comic panels could be a promising direction for future work.

7. REFERENCES

[1] Eakta Jain, Yaser Sheikh, and Jessica Hodgins, "Inferring artistic intention in comic art through viewer gaze," in *Proceedings of the ACM Symposium on Applied Perception*. ACM, 2012, pp. 55–62.

[2] Laurent Itti and Ali Borji, "Computational models of attention," *arXiv preprint arXiv:1510.07182*, 2015.

[3] Ali Borji, Dicky N Sihite, and Laurent Itti, "What stands out in a scene? a study of human explicit saliency judgment," *Vision research*, vol. 91, pp. 62–77, 2013.

[4] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[5] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006, pp. 545–552.

[6] Simone Frintrop, Thomas Werner, and German Martin Garcia, "Traditional saliency reloaded: a good old model in new shape," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 82–90.

[7] Ming Cheng, Niloy J Mitra, Xumin Huang, Philip HS Torr, and Song Hu, “Global contrast based salient region detection,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 3, pp. 569–582, 2015.

[8] Laurent Itti, Christof Koch, and Ernst Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 11, pp. 1254–1259, 1998.

[9] Laurent Itti and Christof Koch, “Computational modelling of visual attention,” *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[10] Ali Borji and Laurent Itti, “State-of-the-art in visual attention modeling,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 185–207, 2013.

[11] Dirk Walther and Christof Koch, “Modeling attention to salient proto-objects,” *Neural networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[12] Laurent Itti and Christof Koch, “A saliency-based search mechanism for overt and covert shifts of visual attention,” *Vision research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[13] Ravi Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk, “Frequency-tuned salient region detection,” in *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*. IEEE, 2009, pp. 1597–1604.

[14] Yun Zhai and Mubarak Shah, “Visual attention detection in video sequences using spatiotemporal cues,” in *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM, 2006, pp. 815–824.

[15] Rajesh PN Rao, Gregory J Zelinsky, Mary M Hayhoe, and Dana H Ballard, “Eye movements in iconic visual search,” *Vision research*, vol. 42, no. 11, pp. 1447–1463, 2002.

[16] Nathan Sprague and Dana Ballard, “Eye movements for reward maximization,” in *Advances in neural information processing systems*. 2003, vol. 16, MIT Press.

[17] Ali Borji, Majid N Ahmadabadi, and Babak N Araabi, “Cost-sensitive learning of top-down modulation for attentional control,” *Machine Vision and Applications*, vol. 22, no. 1, pp. 61–76, 2011.

[18] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu, “Deepfix: A fully convolutional neural network for predicting human eye fixations,” *arXiv preprint arXiv:1510.02927*, 2015.

[19] Matthias Kümmerer, Lucas Theis, and Matthias Bethge, “Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet,” *arXiv preprint arXiv:1411.1045*, 2014.

[20] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao, “Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 262–270.

[21] Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu, “Predicting eye fixations using convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 362–370.

[22] Ali Borji and Laurent Itti, “Cat2000: A large scale fixation dataset for boosting saliency research,” *arXiv preprint arXiv:1505.03581*, 2015.

[23] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba, “Mit saliency benchmark,” http://saliency.mit.edu/.

[24] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch, “Components of bottom-up gaze allocation in natural images,” *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[25] Tilke Judd, Frédo Durand, and Antonio Torralba, “A benchmark of computational models of saliency to predict human fixations,” in *MIT Technical Report*, 2012.

[26] Eero P Simoncelli and William T Freeman, “The steerable pyramid: A flexible architecture for multi-scale derivative computation,” in *icip*. IEEE, 1995, p. 3444.

[27] Paul Viola and Michael Jones, “Robust real-time object detection,” *International Journal of Computer Vision*, vol. 4, 2001.

[28] Aude Oliva and Antonio Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[29] Ruth Rosenholtz, “A simple saliency model predicts a number of motion popout phenomena,” *Vision research*, vol. 39, no. 19, pp. 3157–3163, 1999.