Saliency Prediction for Visual Regions of Interest with Applications in Advertising

Shailee Jain(✉) and S. Sowmya Kamath

National Institute of Technology Karnataka, Surathkal 575025, India
{shaileejain.13it140,sowmyakamath}@nitk.edu.in

Abstract. Human visual fixations play a vital role in a plethora of genres, ranging from advertising design to human-computer interaction. Considering saliency in images thus brings significant merits to Computer Vision tasks dealing with human perception. Several classification models have been developed to incorporate various feature levels and estimate free eye-gazes. However, for real-time applications (Here, real-time applications refer to those that are time, and often resource-constrained, requiring speedy results. It does not imply on-line data analysis), the deep convolution neural networks are either difficult to deploy, given current hardware limitations or the proposed classifiers cannot effectively combine image semantics with low-level attributes. In this paper, we propose a novel neural network approach to predict human fixations, specifically aimed at advertisements. Such analysis significantly impacts the brand value and assists in audience measurement. A dataset containing 400 print ads across 21 successful brands was used to successfully evaluate the effectiveness of advertisements and their associated fixations, based on the proposed saliency prediction model.

Keywords: Visual saliency · Free eye-gaze estimation · Machine Learning · Advertising · Neural networks · Support Vector Machines

1 Introduction

When humans freely view natural scenes, their eyes are drawn to areas that are more prominent amongst background objects. This property, known as *saliency*, refers to the conspicuity of a particular point or region in an image. Free eye-gaze estimation thus aims at determining saliency, with the help of several predictive models. Employed in a plethora of fields, from graphic-oriented tasks (video compression [1], content-aware image resizing [2]), to complex computer vision problems (seam carving [3], non-photorealistic rendering [4]), visual attention prediction plays a pivotal role in determining a user's focus and possibly, manipulating it. This ability to direct a viewer's focus, by incorporating features that play a consequential role in determining saliency, can be extensively applied to use-cases like 'Effective Advertising'.

In this digital era, consumers are constantly being inundated with copious amounts of advertisements. Businesses are resorting to invasive marketing strategies to promote their products aggressively and gain an edge over competitors.

K. Nasrollahi et al. (Eds.): VAAM 2016/FFER 2016, LNCS 10165, pp. 48–60, 2017.
DOI: 10.1007/978-3-319-56687-0_5

A critical part of such strategies, however, lies in audience measurement. When the company is able to thoroughly evaluate the efficacy of its marketing strategy, it is likely to have a stronger impact on potential customers. Measuring the effectiveness of an advertisement is thus an important task in determining the impact of a product. While this can be traditionally done by collecting manual feedback, free eye-gaze estimation enables companies to instead, predict human fixations.

Employing saliency prediction to measure anticipated audience response to advertisements will enable companies to enhance their designs, during the production phase itself. A thorough analysis of human fixations can thus assist the company in re-designing their advertisements, to ensure maximized attention to critical components such as the product, company logo etc. While saliency prediction is an effective tool for such hands-on market analysis, the applicability of existing models is limited. Advanced models such as Deep Convolution Networks are resource/time intensive, while primitive models do not consider multi-level features appropriately, resulting in weak predictions.

Our work, therefore, has two major contributions. Firstly, we analyze and extract various features from an extensive data-set. The resultant tensors are fed into two different classification models - linear kernel Support Vector Machines and a Multi-layer Neural Network. These approaches are then compared to find a suitable classification model, that can sufficiently predict saliency while also being less resource intensive (appropriate for on-line analysis). Secondly, we present a case-study applying the proposed techniques to print advertisements across various genres to find their salient points. The resultant fixations are thoroughly investigated, to evaluate the focus or attraction of an advertisement, and its success. The correlation highlights the applicability of saliency detection in effective design, attention manipulation, and audience measurement through free eye-gaze estimation.

The rest of the paper is organized as follows. Section 2 discusses existing literature in the problem sphere. Section 3 highlights the methodology adopted to develop a saliency model, with specific attention to feature extraction from images and application of two non-linear classifier models. Section 4 presents the experimental results along with an exhaustive case study on the application of saliency models to advertising data, followed by conclusion and references.

2 Related Work

Saliency detection has several primitive hardware oriented approaches involving the use of Eye-trackers. An Eye-tracker is an expensive and cumbersome device that records *saccades* (a rapid movement of the human eye between fixation points) of the user sitting in front of a computer. Given the dynamic nature of applications like audience measurement for advertising and the expansive nature of visual input, device-oriented approaches that consider raw fixation points can be primitive and impractical [5]. Saliency detection models that provide a probabilistic view of prominent locations in an image can instead be used to predict human fixations, even in arbitrary images. Without the limitations of the eye-tracking data-set and sans the need for expensive equipment, saliency prediction can be addressed through alternative computational models. Conventional saliency models are based on 'bottom-up' approach that extracts several intuitive attributes like color contrast, intensity and image center from the picture. Traditional paradigms [6,7] incorporate such low-level, psychologically-backed features to develop a model. In the "Winner-Take-All' and 'Inhibition of Returns' Approach, various features were considered for constructing a linear combination of each map to give the resultant saliency map. The maximum of the result then corresponds to regions of 'highest saliency', to which the visual focus is directed by optical neurons.

In the Itti-Koch model [6], saliency is detected by considering a 2-dimensional layer of integrate-and-fire neurons - multi-layer neurons that 'fire' or alter weights together (global inhibition/reluctance to alter weights). The weights of a neuron activate the inhibition, such that, the first such 'integrate-and-fire' cell to fire is proclaimed the 'winner'. This generates a sequence of action potentials, resulting the FOA (Attention Focus) to shift to the winning location. As a consequence, all cells in the layer are inhibited, to set network in the original state, and find any remaining points of saliency. However, the 'static' scene causes selection of the same 'winner', at all times. Therefore, inhibitory feedback is taken from the winner, such that, within a radius of the FOA, the point and its neighbors are inhibited, allowing different conspicuous locations to be selected, i.e., 'Inhibition-of-Returns'. Nevertheless, these models are limited to images that have a contrast or orientation bias and do not perform well on highly variable human saccades. Also, in spite of a strong center bias that humans hold, as proved by eye-tracking data collected by Judd et al. [8], various other high-level semantic features considered by humans (such as faces or vehicles) are ignored by these approaches. Bottom-up mechanisms, thus, do not completely determine attention selection, as a result of incomplete semantic analysis of the scene.

Context, therefore, remains the most important feature to predict human visual cognition as a more involved activity than merely seeing low-level features. Infusing this into the scene can enhance saliency. For instance, in a textual image like a magazine article, contrast would out play text, unlike a signboard where the text is crucial. Torralba's saliency model [9] effectively incorporates this context in an image. It uses 'Discriminant Spectral Templates' and simple Bayesian classifiers to distinguish scenes as a whole, acting as a precursor to detecting saliency in images, by including semantics. However, it is still incapable of considering high-level features. Thus, the task of feature selection and appropriately combining the same to arrive at a resultant saliency map has a broad scope for research.

The need to understand the scene through a global context, while also extracting primitive information and intuitive features like contrast, and successfully combining the two, to finally determine if a pixel is salient, makes the problem difficult, and virtually impossible with linear classifiers. Recent developments have thus moved onto complex prediction models, that effectively incorporate the entire scene's context and features, to arrive at fixation points. Multiple classification paradigms satisfactorily achieve this task. The most influential are non-linear classifiers, such as Support Vector Machines in Judd's saliency model [5] and Multilayer Neural Networks, coupled with stochastic gradient descent. Judd's model proposes a combination of several low-, mid- and high-level features to holistically analyze contributors to saliency. It eventually achieves high accuracy by combining methods proposed by Itti et al. [6], Hou et al. [7] and Oliva et al. [9]. Jain et al. [10] proved that Judd's model with Torralba's context inclusion doesn't improve accuracy, but effectively intertwines center-bias and provides more advanced features like face detection to capture visual attention.

Recent advancements in the field of Machine Learning and Computer Vision have brought forth a highly complex, and non-intuitive model - a *Convolutional Neural network (CNN)*. CNNs have been proven to provide the best results for all image-related tasks and have been aptly applied to saliency detection. They broadly learn complex features, removing the bias due to human intervention while selecting appropriate input parameters. Nevertheless, it is critical that techniques be adopted for optimizing these time and memory intensive networks, for which effective hardware is still under development, and largely inaccessible.

Although Judd's model has limited classification power, and more advanced models have been developed through deep networks [11,12], it sufficiently includes several feature-ranges to accurately capture saliency. Nevertheless, it is computationally expensive in time to train an SVM with stochastic gradient descent primarily due to its search for appropriate support vectors for margin maximization in the cost function. Also, a linear kernel SVM is essentially a single-layer perceptron (neural network) and a kernelized SVM can be viewed as a Multi-layer Neural Network (MNN) that maximizes the margin in hidden layer space [13]. Bengio et al. [14] proved that given such representations, MNNs learn more intelligently as against shallow architectures such as a linear kernel SVM. In light of these findings, we adopt an approach that successfully incorporated saliency features into an MNN to achieve significant improvement in accuracy when compared to non-linear classifiers like SVM. Our approach is also less resource intensive than deep architectures, and encouragingly effective.

3 Proposed Work

To develop an effective computational model for saliency prediction, non-linear classifiers with multi-level features that appropriately consider intuitive aspects and context of the image, along with high-level semantics were used. These features are fed to a classifier model consisting of SVM and MNNs. The process of determining saliency of given images is described in detail next.

3.1 Features Influencing Saliency

Low-level Features. These features primarily focus on intuitive aspects of the image, largely influenced by its composition. Pixels in an image are made of combinations of primary colors represented. A channel in this context is the gray-scale image of the same size as a color image, made of just one of these primary colors. An RGB image, however, has three channels (red, green, blue); which follow the receptors present in the human eye. In our work, the channels pertaining to these image features are calculated as per Itti-Koch's saliency method. We used the *SaliencyToolbox* [6], a collection of Matlab functions for calculating the saliency map for an image and for serially scanning it as per FOA. Simple features such as orientation, color contrast and intensity are also incorporated in bottom-up saliency. Steerable Pyramid filters in various linear orientations and scales serve as an image decomposition technique to capture energy distribution across the image, influencing its low-level composition.

Mid-level Features. As per the analysis done by Judd et al. and Torralba et al., humans have a natural tendency to look along the horizon as most 'familiar' objects lie on the Earth's surface. Keeping with this, we used a horizon line detector from *LabelMe* [15] as a mid-level feature.

High-level Features. Cognitive visual features such as 'faces', 'humans' and 'vehicles' play a crucial role in saliency and affect human fixations, due to their familiarity and the belief that they hold crucial information while being processed in the user's mind. In early stages of free viewing (first few hundred milliseconds), these image-based conspicuous points guide visual attention. High-level factors like events or actions require a relatively involved analysis by the viewer and thus direct eye movements much later. However, false positive rates significantly impact the performance of saliency models that use such object detectors. These were calculated using the Viola-Jones and Felzenszwalb detectors. Viola-Jones algorithm [16] is a real-time face detection system that uses the image for feature computation, Adaboost for feature selection and an attentional cascade for computational resource allocation. *Felzenszwalb Detector* [17] is a learning based system for detection and localization of objects (vehicles, people) in images, by representing objects using deformable part models.

Center Prior. Humans hold a strong center bias while viewing images. This can be partly attributed to the nature of image capturing wherein the object of interest is more often that not framed near the center of the image. In machine vision, this image center is known as the focus of expansion or the center-of-perspective projection. In our model, focus of expansion is incorporated as a feature indicating the distance of each pixel from the center. Thus, pixels located closer to the center are given more weight/priority, in accordance with the bias.

3.2 Classification Models

There exist several advanced classification models that cannot learn the linear mappings between features and associated class labels. Instead, such models have to be trained to recognize complex non-linear decision boundaries that can appropriately map a feature set with its class. One such involved problem is that of 'Saliency Prediction'. To develop a computational model for the same, we used two non-linear classifiers that were trained on the features obtained after feature extraction. A thorough analysis of the same also proved one approach to be superior to the other (discussed in Sect. 4).

Support Vector Machines. These non-linear classifiers construct a linear decision boundary by mapping complex problems to a higher dimension, where the data behaves linearly. The classifier model focuses on finding the 'Maximum-margin Hyperplane' using 'Support Vectors'. However, due to the involved mathematical computation and resource constraints, 'kernels' are used to reduce complexity. As the saliency data-set is comparatively small when compared to the number of features, the linear kernel is most suited, as against Gaussian or Polynomial kernels that require a huge data-set with much fewer features.

Multi-layer Neural Networks (MNN). A neural network is an advanced non-linear classification model, developed around the structure of our brain. Complex learning tasks are accomplished by a collection of nodes/neurons, connected to each other through weights. The computational task is thus to learn these weight values, by constantly updating them as per the required output. The Network is commonly trained using 'Back-propagation' [18], which carries out optimization using stochastic gradient descent.

3.3 Developing the Saliency Model

The proposed saliency models are described next. The process consists of using a *Feature Extraction* process for obtaining saliency features of data-set images and then using *Classification Models* to generate their saliency maps.

Feature Extraction. The following tools/codes were used for extracting the described low-, mid- and high-level features from the data-set images:

- *Steerable Pyramids* - Used to get the subbands of the steerable pyramids and Torralba saliency model's features [9].
- *Itti and Koch Saliency Toolbox* - Used to get color channels as per the Itti and Koch saliency model.
- *Felzenszwalb and Viola-Jones detectors* - Used to find people, cars and faces in images respectively.
- *LabelMe Toolbox* - Needed for the LMgist.m which is used for the horizon code. The values of the RGB channels, their probabilities as features and the probability of each color as calculated from 3D color histograms are treated as a low-level feature for this model.
- *Distance to center* - indicates the distance from each pixel to center, as these pixels have more significance than pixels farther away from image center.

The second phase involved development of the linear kernel SVM. The input features were fed into the model, to arrive at the required saliency image. Each training image was used to feed 10 positive and 10 negative saliency points from the top 20% salient points and bottom 70% salient points from ground-truth human fixations, respectively. The data-set was divided into 903 training and 100 testing images. The SVM parameter 'C' was empirically found to be 1.

The Multi-layer Neural Network consisted of input nodes that represented the pixels' feature set. It had a single hidden layer and one node, that predicted 1 as salient, and 0 as not (Binary Classification). The model used the same training setup as the SVM and was developed using back-propagation (stochastic gradient descent). Negligible training error was the termination criterion. Weights and bias values were randomly initialised, and optimal value of learning rate was found to be 0.3.

4 Results and Analysis

We used MATLAB R2015a [19] version for Windows (64 bit) for feature extraction and for the development of the proposed saliency model. The MIT dataset[1], consisting of 1003 images depicting natural indoor and outdoor scenes was used for the discussed experiments. A separate advertising data-set comprising of 400 images from 21 different companies across 3 genres was collected for the case study. The SVM was developed using the liblinear library. The Multi-layer Neural Network was designed from ground-up, particularly, for the proposed saliency prediction model.

4.1 Saliency Prediction

Figure 1 shows the saliency maps generated for Neural Network and SVM classifiers. Each map shows a saliency gradient wherein white indicates a highly salient pixel, and black indicates an inconsequential one. The salient area is clearly visible, and from the resultant saliency maps, it is evident that Multi-layer Neural Network (MNN) has performed better than the SVM model. While both capture *actual fixations*[2] to a suitable degree of approximation, the SVM model doesn't always depict the entire salient region and instead focuses on particular points that do not draw as much user attention. The MNN model, alternatively, learns the image structure and accurately predicts saliency, as can be concluded from the similarity between fixations and the MNN saliency map.

The accuracy of both classifiers was compared using the area under ROC curves for various test images. From Fig. 2(a), it can be seen that the MNN model outperforms SVM as is evident from the greater area under the ROC curve. This indicates that the True Positive (TP) Rate increases at a faster rate than the False Positive (FP) Rate, which is of utmost importance in a Computer Vision task. For example, pixels that are wrongly labeled salient could have catastrophic consequences in applications like video compression. Table 1

[1] Available online at http://people.csail.mit.edu/tjudd/WherePeopleLook.

[2] The salient points shown by actual fixations, i.e., raw user inputs (not normalized), depict wagering user attention and thus, do not completely portray salient locations in the image. Instead, they are used as a rough baseline for comparison.

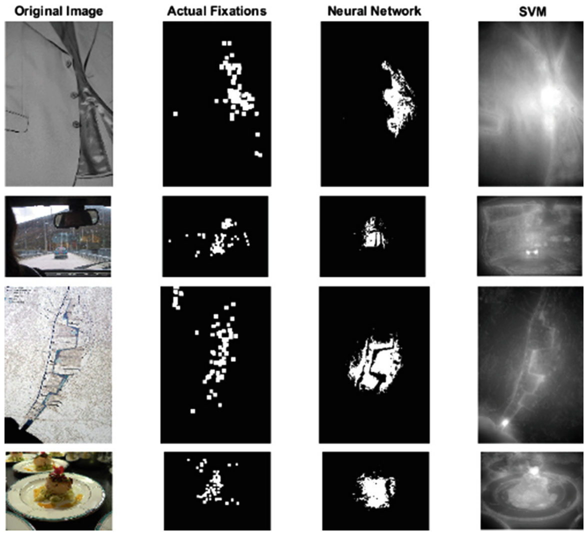
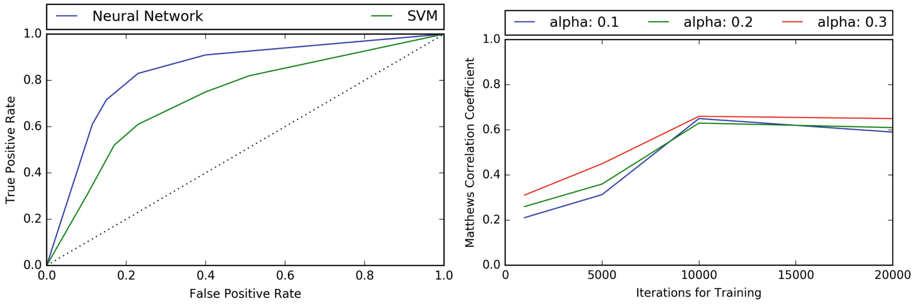


Fig. 1. Saliency Maps generated for MNN and SVM Classifiers

Table 1. Neural network vs. SVM

Method/metric	TPR	FNR	F1 score	Informedness	Markedness
Neural network	0.85	0.36	0.769	0.490	0.513
SVM (Based on [5])	0.74	0.42	0.685	0.320	0.328



(a) ROC for MNN and SVM

(b) MCC for various learning-rates (alpha) and total iterations

Fig. 2. Comparative results of MNN and SVM

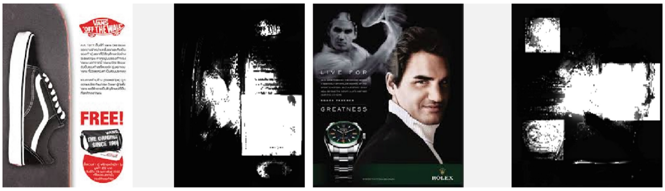
includes the various comparison metrics used for the experimental conditions. The Neural Network is compared against the SVM baseline, on the MIT Saliency Dataset. The previous models indicated in the literature survey include only a subset of these features and use primitive classification techniques, resulting in significantly lower accuracy. Hence, they have been left out of the comparison. Additionally, Matthews Correlation Coefficient (MCC) can be used to gauge effectiveness of binary classification, especially when the class sizes are very different, as in our case (large number of inconsequential points, few salient ones). Figure 2(b) clearly highlights the MCC value for various hyper-parameter values and it can be inferred that appropriate initialization enables high correlation between observation and prediction. Thus, we can conclude that MNN is highly suitable for predicting saliency when compared to the SVM model, while also avoiding the extensive usage of resource and time as is the case of deep networks.

4.2 Case Study: Applying Saliency Prediction in Advertising

Advertisements are the driving force of sales for most of the products that we use today. A substantial amount of money is pumped into ads by companies for marketing and publicity every year. The aim of this case study is to thus find the salient regions in advertisements to check for any correlation between the areas of interest as seen by the consumer (salient regions detected by the model, within acceptable approximation) and positive brand images, for audience measurement. When the regions detected match with the areas of interest, one can presume that the advertisement has the desired effect of publicizing the product. The areas of interest as seen by a majority of consumers generally involves one or more of the following features: *People, Company's Logo and Product, Colorful region or Design.* These features were used for the comparison of areas of interest on the advertisement data-set, which contained print ads of products belonging to different categories like food, lifestyle and electronics. As the proposed MNN saliency detection model performed better in terms of time and accuracy when compared to SVM, it was applied on this data-set for saliency prediction. This map was used as a reference to compare the regions of interest of the advertisement. We present the results for ads for brands in lifestyle (Vans, Rolex), electronics (Nokia, LG) and food (Pepsi, Nestle) here.

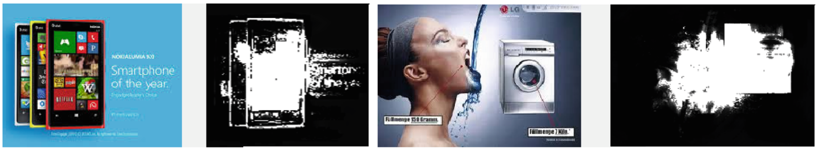
Figure 3 depicts the salient zones obtained for the brand samples for Vans shoes and Rolex watches. As can be seen from the results, the saliency predictions show the print ads satisfy the requirements of an effective advertisement, but can be improved. The company's logo in Fig. 3(a) is rendered salient but is not as eye-catching as the other attributes. For example, the text is presented with the highest saliency, probably in accordance with the company's intention to draw attention to the freebie offer. In the saliency map of Rolex (Fig. 3(b)), the company's brand is the area with the least saliency. Rolex, being a luxury brand, is always associated with celebrities, as can be observed by the dominating saliency region in its print ad. Despite this, the combination of other strong attributes, like a well-placed product picture and the celebrity face significantly boost the efficacy of this print ad. Overall, it can be seen that the above advertisement resonates well with the proposed saliency model, highlighting most key features required to hold user's attention.

In Fig. 4(a), the product (Nokia smartphone) is placed such that it is distinct and demands immediate viewer attention. The ad is not particularly focused on



(a) Vans Shoes (b) Rolex Watches

Fig. 3. Saliency map generation for Lifestyle brands

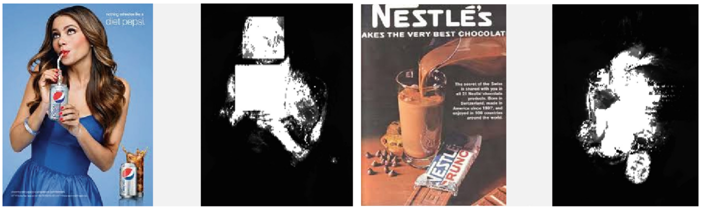


(a) Nokia Smartphones (b) LG Washing Machines

Fig. 4. Saliency map generation for Consumer Electronics brands

the company's logo but the text is saliently depicted, providing useful information that would draw a user's glance and generate interest. Generally, smartphone manufacturers aim at attracting buyers to their product's appearance and their exemplary features in comparison to competitors, which this particular ad achieves successfully. Similarly, for the LG washing machine (Fig. 4(b)), the product itself is depicted reasonably well. The ad is not particularly focused on the company's logo and the text is also not well-defined on the map. However, the use of a person, along with the eye-catching design, captures viewer's attention. If the company's objective was to capture customer attention through its use of bold art, this advertisement is quite effective, as opposed to primal attributes like text and company logo.

In the case of Pepsi (Fig. 5(a)), the product was accurately depicted as a salient zone on the map and the company's logo was also in a well-defined zone. However, the ad text was found to be quite non-salient in this ad, ensuring that user attention is not unnecessarily diverted. The use of a fit person to advertise



(a) Diet Pepsi (b) Nestle

Fig. 5. Saliency map generation for Food brands

the product clearly enhances the saliency in this advertisement, which is for a diet product. Overall, this advertisement responds positively by rendering critical points as salient. In comparison, the Nestle ad (Fig. 5(b)) only the product was placed in a highly salient region, drawing majority of the user's attention. The company's logo and the ad text were not salient or eye-catching. However, as Food brands predominantly focus on their product design and visual appeal to intrigue viewers, this advertisement works well.

Figure 6 shows some results obtained for samples from other ad genres. These results clearly show that the salient regions obtained from advertisements have a strong correlation with the areas of interest. These reputed, brand-oriented companies are thus strongly linked to 'effective' advertisements[3] that further propagate their marketing. Such techniques of audience measurement can effectively be used by companies to evaluate their designs in early stages of development, enabling them to appeal to their potential customers better. Improvements can also be done using saliency features, such as bold design on known faces. Consequentially, effective advertisements designed after thorough investigation of probable impact will boost sales and maximize product outreach.

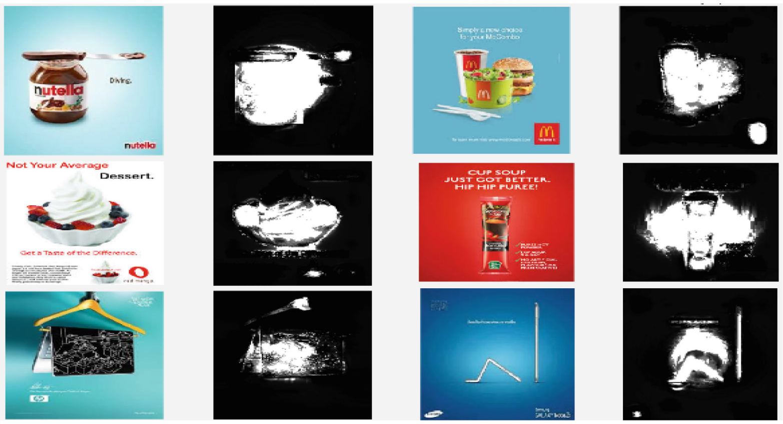


Fig. 6. Saliency Maps for sample ads belonging to other genres

5 Conclusion and Future Work

In this paper, a novel Multi-layer Neural Network approach for the Computer Vision task of saliency prediction was presented. This model comprises of a non-linear classifier that incorporates multilevel features like intuitive attributes (like contrast), image semantics and face detection. The proposed classifier model is

[3] By effectiveness, we mean that the advertisement highlights the product, company etc. and immediately captures consumer attention, sparking interest.

extensively trained, tested and compared to human fixation points. Its effectiveness was also compared with that of a SVM model, in terms of prediction accuracy. The obtained experimental results highlight the success of our approach, while also depicting significant improvements when compared to the SVM. The MNN model also has lower resource usage (does not require specific hardware and has reasonable training time), making it suitable for real-time tasks with time, resource constraints. A case-study to demonstrate the role of saliency in effective advertising and its contribution to successful product/brand images was also presented. Using the proposed MNN saliency prediction model, it was found that there exists a strong correlation between the outreach of an advertisement and its ability to render important attributes salient, such as the product or company name. Advertisement applications can, therefore, be developed significantly along the lines of saliency prediction, for effective marketing.

As part of future work, we intend to explore the possibility of the existence of a correlation between the features incorporated in our saliency model *(hand-engineered)* and those extracted by a Convolution Neural Network *(found using reverse engineering)*. We also intend to extend this study to deep networks on development of appropriate, feasible architectures to analyze further significant improvements in real-time (time constrained) applications like Advertising. This can be used as an alternative to our saliency model while carrying out an extended study to tasks that do not require speedy analysis but need to incorporate a learnable model.

References

1. Wang, Z., Lu, L., Bovik, A.C.: Foveation scalable video coding with automatic fixation selection. IEEE Trans. Image Process. **12**(2), 243–254 (2003)
2. Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., Cohen, M.: Gaze-based interaction for semi-automatic photo cropping, pp. 771–780 (2006)
3. Rubinstein, M., Shamir, A., Avidan, S.: Improved seam carving for video retargeting. ACM Trans. Graph. **27**(3) 16:1–16:9 (2008)
4. DeCarlo, D., Santella, A.: Stylization and abstraction of photographs. ACM Trans. Graph. **21**(3), 769–776 (2002)
5. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. 2106–2113 (2009)
6. Itti, L., Koch, C., Niebur, E., et al.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
7. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach, pp. 1–8 (2007)
8. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations (2012)
9. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
10. Jain, E., Mukerjee, A., Kochhar, S.: Predicting where humans look in a visual search: Incorporating context-based guidance
11. Borji, A., Tavakoli, H.R., Sihite, D.N., Itti, L.: Analysis of scores, datasets, and models in visual saliency prediction. In: IEEE ICCV 2013, pp. 921–928. IEEE (2013)

12. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: IEEE Conference of Computer Vision and Pattern Recognition (CVPR), pp. 1265–1274 (2015)
13. Collobert, R., Bengio, S.: Links between perceptrons, MLPs and SVMs. In: Proceedings of the Twenty-First International Conference on Machine Learning, ICML 2004, p. 23. ACM, New York (2004)
14. Bengio, Y., LeCun, Y., et al.: Scaling learning algorithms towards AI. Large-scale Kernel Mach. **34**(5), 1–41 (2007)
15. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. Int. J. Comput. Vis. **77**(1), 157–173 (2008)
16. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
17. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model, pp. 1–8 (2008)
18. Bottou, L., Cortes, C., Denker, J.S., Drucker, H., Guyon, I., Jackel, L.D., LeCun, Y., Muller, U.A., Sackinger, E., Simard, P., et al.: Comparison of classifier methods: a case study in handwritten digit recognition. In: ICPR, pp. 77–87 (1994)
19. The Mathworks, Inc. Natick, Massachusetts: MATLAB version 8.5.0.197613 (R2015a) (2015)