The AAAI 2017 Spring Symposium on
Science of Intelligence: Computational Principles of Natural and Artificial Intelligence
Technical Report SS-17-07

Predicting When Eye Fixations Are Consistent

Anna Volokitin,[1] **Michael Gygli,**[1] **Xavier Boix**[2]
[1] Computer Vision Laboratory, ETH Zurich, Switzerland
[2] CBMM, Massachusetts Institute of Technology & Istituto Italiano di Tecnologia, Cambridge, MA (USA)

Abstract

Many computational models of visual attention use image features and machine learning techniques to predict eye fixation locations as saliency maps. Recently, the success of Deep Convolutional Neural Networks (DCNNs) for object recognition has opened a new avenue for computational models of visual attention due to the tight link between visual attention and object recognition. In this paper, we show that using features from DCNNs for object recognition we can make predictions that enrich the information provided by saliency models. Namely, the consistency of the eye fixations among subjects, *i.e.* the agreement between the eye fixation locations of different subjects, can be predicted.

Introduction

This paper is a shortened version of our earlier work (Volokitin, Gygli, and Boix 2016).

Gaze shifting allocates computational resources by selecting a subset of the visual input to be processed, *c.f.* (Ungerleider 2000). Computational models of visual attention provide a reductionist view on the principles guiding attention. These models are used both to articulate new hypotheses and to challenge the existing ones. Machine learning techniques that can make predictions directly from the image have facilitated the study of visual attention in natural images. Also, these models have found numerous applications in visual design, image compression, and some computer vision tasks such as object tracking.

Many computational models of attention predict the image location of eye fixations, which is represented with the so called saliency map. The seminal paper by Koch and Ullman introduced the first computational model for saliency prediction (Koch and Ullman 1985). This model is rooted in the feature integration theory, that pioneered the characterisation of many of the behavioural and physiological observed phenomena of visual attention (Treisman and Gelade 1980). Since then, a rich variety of models have been introduced to extract the saliency map, *e.g.* (Harel, Koch, and Perona 2007; Itti, Koch, and Niebur 1998; Judd et al. 2009; Kienzle et al. 2006; Walther and Koch 2006).

Some authors stressed the need to predict properties of the eye fixations beyond the saliency map to study different phenomena of visual attention and to allow for new applications, *e.g.* (Jiang et al. 2015; Le Meur, Baccino, and Roumy 2011; Mathe and Sminchisescu 2013). Since visual attention is strongly linked to object recognition, the advent of near-human performing object recognition techniques based on DCNNs opens a new set of possibilities for models of visual attention. In this paper, we analyze two ways to augment the eye fixation location information delivered by saliency models by using features extracted from DCNNs trained for object recognition.

We show that the consistency of eye fixation locations among subjects can be predicted from features based on object recognition. In Fig. 1 we show images with different degrees of consistency among subjects, that illustrate that eye fixation consistency varies depending on the image. There is a plethora of results in the literature showing that consistency varies depending on the group the subjects belong to. There are marked differences between subjects with autism spectrum disorders and those without (Dalton et al. 2005; Klin et al. 2002), between subjects from different cultures (Chua, Boland, and Nisbett 2005), and between fast and slow readers (Kliegl, Nuthmann, and Engbert 2006). Yet, the causes of eye fixation inconsistencies among individual subjects rather than for groups may be difficult to explain in natural images, especially because natural images are not designed to isolate a specific effect.

The model we introduce to predict the eye fixation consistency substantially improves the performance of a previous attempt (Le Meur, Baccino, and Roumy 2011), and it shows that the eye fixation consistency depends on the object categories present in the image.

Finally, our results reveal that, like memorability (Isola et al. 2011) and interestingness (Gygli et al. 2013), eye fixation consistency is an attribute of natural images that can be predicted.

Predicting the Eye Fixations Consistency

Datasets We use the MIT (Judd et al. 2009) dataset, which includes 1003 images with everyday indoor and outdoor scenes. All images are presented to 15 subjects for 3 seconds. This dataset is a standard benchmark to evaluate the prediction of eye fixation locations in natural images.

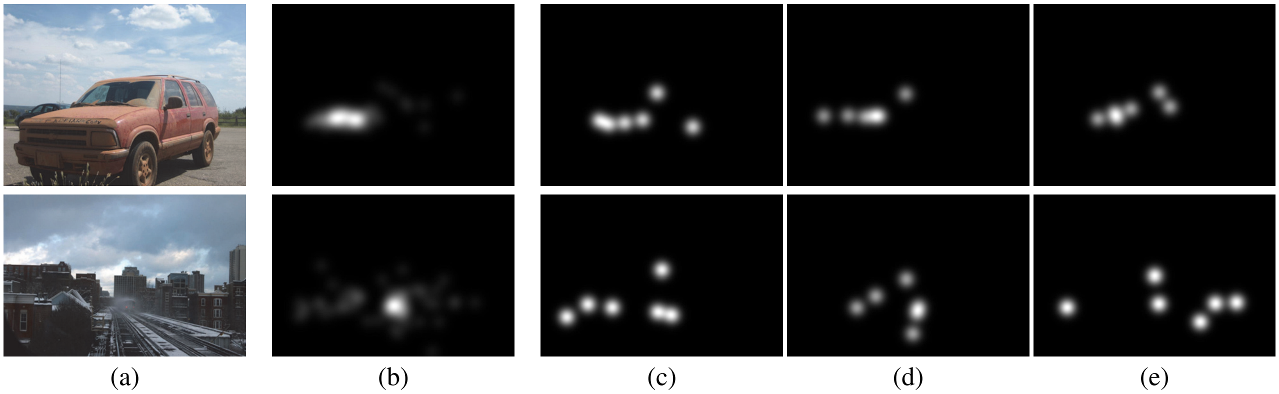


Figure 1: Fixations from individual subjects. (a) the raw image, (b) averaged fixation map, (c) - (e) individual fixations from subjects. The top row shows an image where fixations are highly consistent, and the bottom shows one where the fixations are inconsistent.

Eye Fixation Maps An eye fixation map is constructed for each subject by taking the set of locations where the eyes are fixated for a certain period of time (conventionally taken to be 50ms). The fixation map is a probability distribution over salient locations in an image, and ideally would be computed by taking an average over infinite subjects. In practice, the eye fixation map is computed by summing eye fixation maps of the individual subjects (which are binary images, with ones at fixation locations and zeroes elsewhere). The result is smoothed with a Gaussian of width dependent on the eye tracking set up (1 degree of visual angle in the MIT dataset). Finally, the map is normalised to sum to one.

To study viewing patterns in an image, we estimate the eye fixation consistency among subjects, *i.e.* the amount of inter-subject variability in viewing the image. To do this, we first measure the true eye fixation consistency given the eye fixations of individual subjects, adapting a procedure used in (Torralba et al. 2006), which we introduce next.

Metric of the Eye Fixation Consistency The eye fixation consistency metric tests whether the fixation map computed from a subset of subjects can predict the fixation map computed from the rest of the subjects. Let $\mathcal{O}$ be the set of all subjects (*e.g.* 15 in MIT dataset), and $\mathcal{H}$ be the subset of K subjects held out for testing. We compute two eye fixation maps: $M_{\mathcal{H}}$ from $\mathcal{H}$, and $M_{\mathcal{O}\setminus\mathcal{H}}$ from $\mathcal{O} \setminus \mathcal{H}$ (the remaining $15 - K$ subjects). We define the consistency score to be the score of $M_{\mathcal{H}}$ in predicting $M_{\mathcal{O}\setminus\mathcal{H}}$ using any of the standard metrics for evaluating saliency prediction algorithms (introduced in the section below). To be consistent in our evaluation of consistency, $M_{\mathcal{H}}$ is treated as the saliency map, and $M_{\mathcal{O}\setminus\mathcal{H}}$ as the eye fixation map, as it is computed from more subjects than $M_{\mathcal{H}}$. We set $K = 7$.

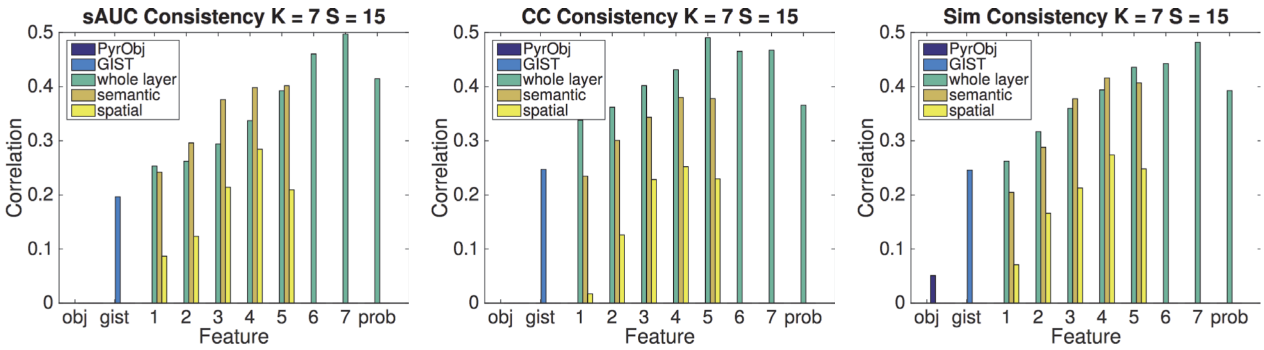
Metric of the Saliency Map Accuracy Since there is no consensus among researchers about which metric best captures the accuracy of the saliency map (*c.f.* (Riche et al. 2013)), we follow the lead of (Judd, Durand, and Torralba 2012) and report 3 metrics. Now we briefly define the metrics used in this paper, and refer the reader to (Riche et al. 2013) for a more complete treatment. Under all of these metrics a higher score indicates better performance. Below, M_F is the map of eye fixation map (ground truth) and M_S is the (predicted) saliency map:

- *Similarity (Sim).* The similarity metric is also known as the histogram intersection metric, and it is defined as $S = \sum_x \min(M_F(x), M_S(x))$.
- *Cross Correlation (CC).* This metric quantifies to what extent there is a linear relationship between the two maps. It is defined: $CC = cov(M_F, M_S)/(\sigma_{M_F}\sigma_{M_S})$, where σ_M is the standard deviation of the map M.
-*Shuffled Area under the Curve (sAUC).* The saliency map is treated as a binary classifier to separate positive from negative samples at various intensity thresholds. It is called shuffled because the points of the saliency map are sampled from fixations on other images to discount the effect of center bias. This metric can take values between 0.5 and 1. Although the previous two metrics are symmetric, meaning the two maps are interchangeable, this one is not.

Computational Model

To predict consistency we train a regressor between the features extracted from the image and the response variable. The features and learner are the same same for both applications. We use a Support Vector Regressor (Vapnik 1995) with the χ^2 kernel. We introduce several image features to test the hypothesis that consistency can be predicted from the spatial distribution and the categories of the objects in the image. The splits are done taking randomly 60% of images for training and the rest for testing. The learning parameters are set with a 10 fold cross-validation using LIBSVM to determine the cost C (range 2^{-4} to 2^{6}) and ϵ (2^{-8} to 2^{-1}) of the ϵ-SVR.

Deep Convolutional Neural Networks To capture the spatial distribution and category of the objects in the image, we use features taken from the layers of a DCNN. A DCNN is a feedforward neural network with constrained



(a) (b) (c)

Figure 2: Evaluation of the Prediction of the Eye Fixations Consistency. The correlation between the predicted consistency of the eye fixation and the true consistency is evaluated using different input features (including each of the 7 layers of the DCNN). The metric used to evaluate the consistency uses $K = 7$ subjects and is averaged over $S = 15$ random splits and is based on: (a) sAUC, (b) CC, and (c) Sim. The results show a similar trend with different values of K

connections between layers, that take the form of convolutions or spatial pooling, besides other possible non-linearities, *e.g.* (LeCun et al. 1990; Krizhevsky, Sutskever, and Hinton 2012; Pinto et al. 2009). We use the DCNN called AlexNet (Krizhevsky, Sutskever, and Hinton 2012) with trained parameters in ImageNet, which achieved striking results in the task of object recognition. It consists of eight layers, the last three of which are fully connected.

Let $\mathbf{y}_l$ be a two-dimensional matrix that contains the responses of the neurons of the DCNN at the layer l. $\mathbf{y}_l$ has size $s_l \times d_l$, that varies depending on the layer. The first dimension of the table indexes the spatial location of the center of the neuron's receptive field, and the second dimension indexes the patterns to which the neuron is tuned. The response of a neuron, $y_l[i][j]$, has a high response when pattern j is present at location i. Neural responses at higher layers in the network encode more meaningful semantic representations than at lower layers (Zeiler and Fergus 2014), but the spatial resolution at the last layers is lower than at the first layers.

The neural responses from the top of each layer $\mathbf{y}_l$ are used as features.

Spatial Distribution of Objects We introduce two different features to capture the spatial distribution of the objects without describing their object categories. The first feature is based on the DCNNs previously introduced. We take the neural responses in a layer, $\mathbf{y}_l$, and convert them into a feature that has one response for each location that corresponds to the presence of a pattern or object detected by the CNN (it has dimensions $s_l \times 1$). To do so, we discard information about which pattern is present at a certain location and simply take the highest response among the patterns. Thus, the image feature is $f_l[i] = \max_j y_l[i][j]$. This corresponds to max pooling over the pattern responses.

A second feature we introduce is based on the objectness, or the likelihood that a region of an image contains an object of any class (Alexe, Deselaers, and Ferrari 2012). Objectness is based on detecting properties that are general for any object, such as the closedness of boundaries. We use the code provided by (Cheng et al. 2014) to generate bounding boxes ranked by the probability that they contain an object. We take the top 500 boxes to create a heatmap. The intensity of each pixel in this heatmap is proportional to the number of times it has been included in an objectness proposal We divide the heatmap into sub-regions at four different levels of resolution and evaluate the L_2 energy in each sub-region, creating a spatial pyramid (Lazebnik, Schmid, and Ponce 2006). This feature gives an indication of how objects are located in the image. We call this feature PyrObj.

Object Categories For each not fully connected layer of the DCNN, we construct a feature with only semantic information analogously to the feature with only spatial information. This image feature is $f_l[j] = \max_i y_l[i][j]$, and is of dimension $1 \times d_l$. This corresponds to max pooling over space. The last layers of the DCNN already capture object categories, as they transform the neural responses to object classification scores that contain little to no information about the location of the objects in the image.

Gist of the scene This descriptor of length 512, introduced by (Oliva and Torralba 2001), gives a representation of the structure of real world scenes where local object information is discarded. Scenes belonging to the same semantic categories (such as streets, highways and coasts) have similar GIST descriptors.

Predicting the Eye Fixations Consistency

We now evaluate the performance of the prediction of the eye fixation consistency. We report the Spearman correlation between the true and predicted values in Fig. 2.

The results show that the PyrObj objectness feature can partially describe the object distribution and performs similarly to the spatial features of the DCNN. In general, Gist performs better than PyrObj, on par with the best spatial feature. Interestingly, we see that the semantic feature is much more informative for predicting consistency than the spatial feature, which suggests that semantic information has a greater contribution to predicting consistency than information about the distribution of the objects. Subsequent layers outperform the preceding ones, except of the last *prob* layer, which performs slightly worse. This could happen because the *prob* layer has lost all spatial information. Finally, note that the best performing feature is the whole layer of the DCNN, achieving a ρ of around 0.5.

The previous work that also used machine learning to predict the eye fixation consistency (Le Meur, Baccino, and Roumy 2011), reports a Pearson correlation of 0.27 on a set of 27 images they have selected at hand, which shows the challenge of this task. Our results substantially improve over previous work, mainly because we use features based on object recognition. Our results reveal that the eye fixation consistency among subjects is an attribute of natural images that can be predicted.

Conclusions

We used machine learning techniques and automatic feature extraction to predict the eye fixation consistency among subjects in natural images. This was possible due to the good performance of DCNNs for object recognition, since eye fixations locations are strongly related to the object categories. Our results showed that the eye fixation consistency among subjects is an attribute of natural images that can be predicted from object categories. We expect that all these results allow for numerous applications in computer vision and visual design.

Acknowledgements We thank Luc Van Gool and Qi Zhao for their useful comments and advice. This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 2 under Grant R-263-000-B32-112, the Singapore Defence Innovative Research Programme under Grant 9014100596, and the ERC Advanced Grant VarCity.

References

Alexe, B.; Deselaers, T.; and Ferrari, V. 2012. Measuring the objectness of image windows. *TPAMI.*

Cheng, M.-M.; Zhang, Z.; Lin, W.-Y.; and Torr, P. 2014. Bing: Binarized normed gradients for objectness estimation at 300fps. In *CVPR.*

Chua, H. F.; Boland, J. E.; and Nisbett, R. E. 2005. Cultural variation in eye movements during scene perception. *PNAS.*

Dalton, K. M.; Nacewicz, B. M.; Johnstone, T.; Schaefer, H. S.; Gernsbacher, M. A.; Goldsmith, H.; Alexander, A. L.; and Davidson, R. J. 2005. Gaze fixation and the neural circuitry of face processing in autism. *Nature neuroscience*.

Gygli, M.; Grabner, H.; Riemenschneider, H.; Nater, F.; and Gool, L. V. 2013. The interestingness of images. In *ICCV*.

Harel, J.; Koch, C.; and Perona, P. 2007. Graph-based visual saliency. In *NIPS*.

Isola, P.; Xiao, J.; Torralba, A.; and Oliva, A. 2011. What makes an image memorable? In *CVPR*.

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* 20(11):1254–1259.

Jiang, M.; Boix, X.; Roig, G.; Xu, J.; Van Gool, L.; and Zhao, Q. 2015. Learning to predict sequences human visual fixations. *TNNLS*.

Judd, T.; Ehinger, K.; Durand, F.; and Torralba, A. 2009. Learning to predict where humans look. In *ICCV*.

Judd, T.; Durand, F.; and Torralba, A. 2012. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*.

Kienzle, W.; Wichmann, F.; Schölkopf, B.; and Franz, M. 2006. A nonparametric approach to bottom-up visual saliency. In *NIPS*.

Kliegl, R.; Nuthmann, A.; and Engbert, R. 2006. Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of experimental psychology: General*.

Klin, A.; Jones, W.; Schultz, R.; Volkmar, F.; and Cohen, D. 2002. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of general psychiatry*.

Koch, C., and Ullman, S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*.

Le Meur, O.; Baccino, T.; and Roumy, A. 2011. Prediction of the inter-observer visual congruency (iovc) and application to image ranking. In *ACM international conference on Multimedia*.

LeCun, Y.; Boser, B. E.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W. E.; and Jackel, L. D. 1990. Handwritten digit recognition with a back-propagation network. In Touretzky, D. S., ed., *Advances in Neural Information Processing Systems 2*. Morgan-Kaufmann. 396–404.

Mathe, S., and Sminchisescu, C. 2013. Action from Still Images Datasets and Models to Learn Task Specific Human Visual Scanpaths. In *NIPS*.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*.

Pinto, N.; Doukhan, D.; DiCarlo, J. J.; and Cox, D. D. 2009. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computational biology*.

Riche, N.; Duvinage, M.; Mancas, M.; Gosselin, B.; and Dutoit, T. 2013. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *ICCV*.

Torralba, A.; Oliva, A.; Castelhano, M. S.; and Henderson, J. M. 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*.

Treisman, A., and Gelade, G. 1980. A feature-integration theory of attention. *Cognitive Psychology*.

Ungerleider, S. 2000. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*.

Vapnik, V. 1995. *The nature of statistical learning theory*. Springer Science & Business Media.

Volokitin, A.; Gygli, M.; and Boix, X. 2016. Predicting when saliency maps are accurate and eye fixations consistent. In *CVPR*.

Walther, D., and Koch, C. 2006. Modeling attention to salient proto-objects. *Neural Networks*.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *ECCV*.