Decoding What People See from Where They Look: Predicting Visual Stimuli from Scanpaths

Moran Cerf[1,⋆,⋆⋆], Jonathan Harel[1,⋆], Alex Huth[1], Wolfgang Einhäuser[2], and Christof Koch[1]

[1] California Institute of Technology, Pasadena, CA, USA
moran@klab.caltech.edu
[2] Philipps-University Marburg, Germany

Abstract. Saliency algorithms are applied to correlate with the overt attentional shifts, corresponding to eye movements, made by observers viewing an image. In this study, we investigated if saliency maps could be used to predict which image observers were viewing given only scanpath data. The results were strong: in an experiment with 441 trials, each consisting of 2 images with scanpath data - pooled over 9 subjects - belonging to one unknown image in the set, in 304 trials (69%) the correct image was selected, a fraction significantly above chance, but much lower than the correctness rate achieved using scanpaths from individual subjects, which was 82.4%. This leads us to propose a new metric for quantifying the importance of saliency map features, based on discriminability between images, as well as a new method for comparing present saliency map efficacy metrics. This has potential application for other kinds of predictions, e.g., categories of image content, or even subject class.

1 Introduction

In electrophysiological studies, the ultimate validation of the relationship between physiology and behavior is the decoding of behavior from physiological data alone [1,2,3,4,5,6,7]. If one can determine which image an observer has seen using only the firing rate of a single neuron, one can conclude that that neuron's output is highly informative about the image set. In psychophysical studies it is common to show an observer (animal or human) a sequence of images or video while recording their eye movements using an eye-tracker. Often, such studies aim to predict subjects' scanpaths using saliency maps [8,9,10,11], or other techniques [12,13]. The predictive power of a saliency model is typically judged by computing some similarity metric between scanpaths and the saliency map generated by the model [8,14]. Several similarity metrics have become de facto standards, including NSS [15] and ROC [16]. A principled way to assess the goodness of such a metric is to compare its value for scanpath-saliency map pairs which correspond to the same image and different images. If this difference

[⋆] These authors contributed equally to this work.

[⋆⋆] Corresponding author.

L. Paletta and J.K. Tsotsos (Eds.): WAPCV 2008, LNAI 5395, pp. 15–26, 2009.

is systematic, one can apply the metric to several candidate saliency maps per image, and asses which saliency map yields the highest decodability.

This decodability represents a new measure of saliency map efficacy. It is complementary to the current approaches: rather than predicting fixations from image statistics, it predicts image content from fixation statistics. The fundamental advantage of rating saliency maps in this way is that the score reflects not only how similar the scanpath is to the map, but also how *dissimilar it is from the maps of other images.* Without that comparison, it is possible to artificially inflate similarity metrics using saliency heuristics which increase the correlation with all scanpaths, rather than only those recorded on the corresponding image. Thus, we propose this as an alternative to the present measures of saliency maps' predictive power, and test this on established eye-tracking datasets.

The contributions of this study are:

1. A novel method for quantifying the goodness of an attention prediction model based on the stimuli presented and the behavior.
2. Quantitative results using this method that rank the importance of feature maps based on their contribution to the prediction.

2 Methods

2.1 Experimental Setup

In order to test if scanpaths could be used to predict which image from a set was being observed at the time it was recorded, we collected a large dataset of images and scanpaths from various earlier experiments (from the database of [17]). In all of these previous experiments, images were presented to subjects for 2 s, after which they were instructed to answer "How interesting was the image?" on a scale of 1-9 (9 being the most interesting). Subjects were not instructed to look at anything in particular; their only task was to rate the entire image. Subjects were always naïve to the purpose of the experiments. The subset of images was presented for each subject in random order.

Scenes were indoors and outdoors still images (see examples in Fig. 1), containing faces and objects. Faces were in various skin colors and age groups, and exhibiting neutral expressions. The images were specifically composed so that the faces and objects appeared in a variety of locations but never in the center of the image, as this was the location of the starting fixation on each image. Faces and objects vary in size. The average size was $5\% \pm 1\%$ (mean ± s.d.) of the entire image - between 1° to 5° of the visual field. The number of faces in the images was varied between 1-6, with a mean of 1.1 ± 0.48 (s.d.). 441 images (1024×768 pixels) were used in these experiments altogether. Of these, 291 images were unique. The remaining 150 stimuli consisted of 50 different images that were repeated twice, but treated uniquely as they were recorded under different experimental conditions. Of the unique images, some were very similar to each other, as only foreground objects but not the background was changed. Since we only counted finding the exact same instance (*i.e.* 1 out of 441) as correct

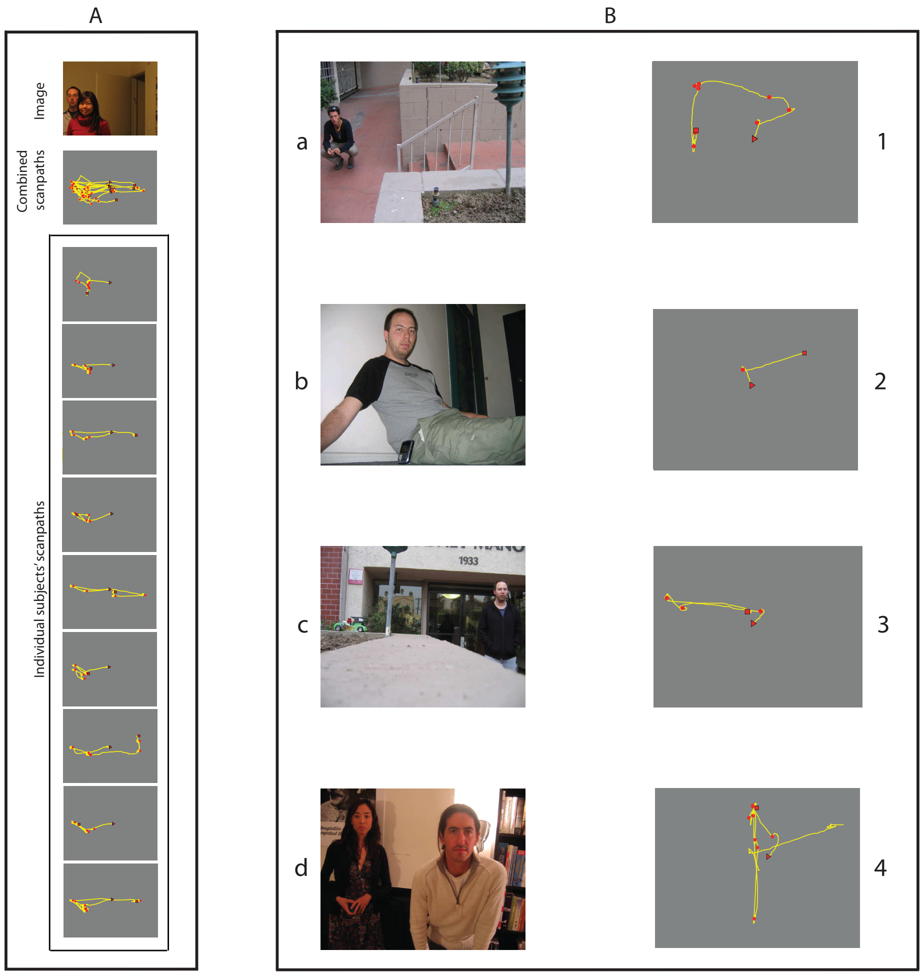

Fig. 1. Examples of scanpaths/stimuli used in the experiment. **A.** Scanpaths of the 9 individual subjects used in the analysis for a given image. The combined fixations of all subjects was used for further analysis of the agreement across all subjects, and for analysis of the ideal subjects' pool size for decoding. The red triangle marks the first and the red square the last fixation, the yellow line the scanpath, and the red circles the subsequent fixations. Top: the image viewed by subjects to generate these scanpaths. The trend of visiting the faces – a highly attractive feature – yields greater decoding performance. **B.** Four example images from the dataset (left) and their corresponding scanpaths for different arbitrary chosen individuals (right). Order is shuffled. See if you can match ("decode") the scanpath to its corresponding images. *The correct answers are: a3, b4, c2 and d1.*

prediction, in at least $\frac{150}{441} \times \frac{2}{440} = 0.15\%$ of cases a nearly correct prediction (same or very similar image) would be counted as incorrect. Hence, our datasets are challenging and the estimates of correct prediction conservative.

Eye-position data were acquired at 1000 Hz using an Eyelink1000 (SR Research, Osgoode, Canada) eye-tracking device. The images were presented on a CRT2 screen (120 Hz), using MATLAB's Psychophysics and eyelink toolbox extensions. Stimulus luminance was linear in pixel values. The distance between the screen and the subject was 80 cm, giving a total visual angle for each image of $28° \times 21°$. Subjects used a chin-rest to stabilize their head. Data were acquired from the right eye alone. Data from a total of nine subjects, each with normal or corrected-to-normal vision, were used. We discard the first fixation from each scanpath to avoid adding trivial information from the initial center fixation. Thus, we worked with $441 \times 9 = 3969$ total scanpaths.

2.2 Decoding Metric

For each image, we created six different "feature maps". Four of the maps were generated using the Itti and Koch saliency map model [8]: (1) combined color-intensity-orientation (CIO) map, (2) color alone (C), (3) intensity alone (I), and (4) orientation alone (O). A "faces" map was generated using the Viola and Jones face recognition algorithm [18]. The sixth map, which we call "CIO+F" was a combination of the face map and the CIO map from the Itti and Koch saliency model, which has been shown to be more predictive of observers fixations than CIO [17]. Each feature map was represented as a positive valued heat map over the image plane, and downsampled substantially, in line with [8], in our case to nine by twelve pixels, each pixel corresponding to roughly 2×2 degrees of visual angle. Subject fixation data was binned into an array of the same size. The saliency maps and fixation data were compared using an ROC-based method [16]. This method compares saliency at fixated and non-fixated locations (see Fig. 2 for an illustration of the method). We assume some threshold saliency level above which locations on the saliency map are considered to be predictions of fixation. If there is a fixation at such a location, we consider it a hit, or true positive. If there is no fixation, it is considered a false positive. We record the true positive and false positive rates as we vary the threshold level from the minimum to the maximum value of the saliency map. Plotting false positive vs. true positive results in a Receiver Operator Characteristics ("ROC") curve. We integrate the Area Under this ROC Curve ("AUC") to get a scalar similiarity measure (AUC of 1 indicates all fixations fall on salient locations, and AUC of 0.5 is chance level). The AUC for the correct scanpath-image pair was ranked against other scanpath-image pairs (from 1 to 31 decoy images, chosen randomly from the remaining 440 to 410 images), and the decoding was considered successful only if the correct image was ranked one. In the largest image set size we tried, if any of the other 31 AUCs for scanpath/images was higher than the one of the correct match, we considered the prediction a miss (e.g. for one decoding trial the algorithm would be as follows: *1.* Randomly select a scanpath out of the 3969 scanpaths. *2.* Consider the image it belongs to, together with 1 to 31 randomly selected decoys. We will attempt to match the scanpath to its associated image out of this set of candidates. *3.* Compute a feature map for each image in the candidate set. *4.* Compute the AUC of the scanpath for each of the 2-32 saliency

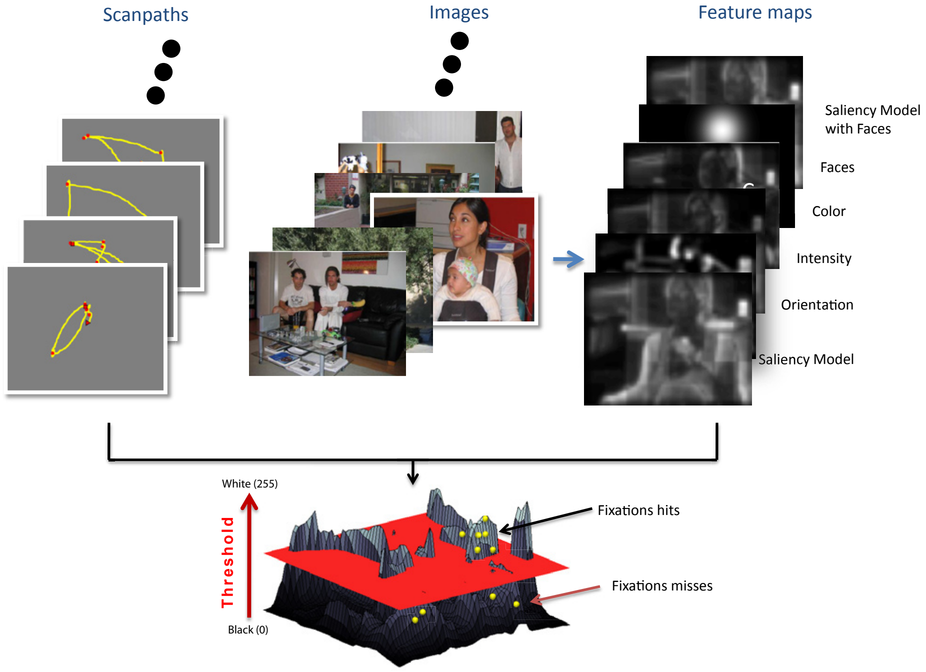

Fig. 2. Illustration of the AUC calculation. For each scanpath, we choose the corresponding image and 1–31 decoys. For each image we calculate each of the 6 feature maps (C, I, O, F, CIO, CIO+F). For a given scanpath and a feature map we then calculate the ROC by varying a threshold over the feature plane and counting how many fixations fall above/below the threshold. The area under the ROC curve (AUC) serves as a measure of agreement between the scanpath and the feature map. We then rank the images by their AUC scores, and consider the decoding correct if the highest AUC is that of the correct image.

maps. *5.* Decoding is considered successful iff the image on which the scanpath was actually recorded has the highest AUC score.).

3 Results

We calculated the average success rate of prediction trials, each of which consists of (1) fixations pooled over 9 subjects' scanpaths, and (2) an image set of particular cardinality, from 2 to 32, ranked according to the ROC-fixation score on one of three possible feature maps: CIO, CIO+F, or F. We used the face channel although it carries some false identifications of faces, and some misses, as it has been shown to have higher predictive power, involving high-level (semantic) saliency content with bottom-up driven features [17]. We reasoned that using the face channel alone in this discriminability experiment would provide a novel method of comparing it to saliency maps' predictive power.

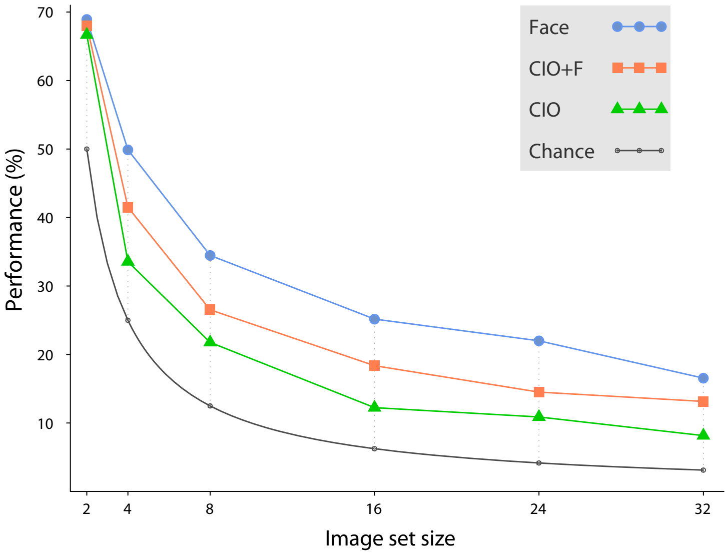

Fig. 3. Decoding performance with respect to image pool size. Decoding with scanpaths pooled over 9 subjects, we varied the number of decoy images used between 1 and 31. The larger the image set size, the more difficult the decoding. For each image set size and scanpath we calculated the ROC over 3 feature maps: a face-channel which is the output of the Viola and Jones face-detection algorithm with the given image (F), a saliency map based on the color, orientation and intensity maps (CIO), and a saliency map combining the face-channel and the color, orientation and intensity maps (CIO+F). While all feature maps yielded a similar decoding performance for the smaller pool size, the performance was least degraded for the F map. The face feature map is higher than the CIO+F map and the two are higher than the CIO map. All maps predict above chance level – shown in the bottom line as the multiplicative inverse of the image set size.

For one decoy per image set (image set size = two), we find that the face feature map (F) was used to correctly predict the image seen by the subjects in 69% of the trials ($p < 10^{-15}$, sign test[1]), while the CIO+F feature map was correct in 68% ($p < 10^{-14}$), and CIO in 66% ($p < 10^{-12}$) of trials. This $F > CIO + F > CIO$ trend persists through all image set sizes. Pooling prediction trials over all image set sizes (6 sizes × 441 trials per size = 2646 trials), we find that using the F map yields a prediction that is at least as accurate as the CIO map in 89.9% of trials, with significance $p < 10^{-8}$ using the sign-test. Similarly, F is at least as predictive as CIO+F in 90.3% of trials ($p < 10^{-15}$), and CIO+F is at least as predictive as CIO in 97.8% of trials ($p < 10^{-21}$). All data points

[1] The sign-test tests against the null hypothesis that the distribution of correct decodings is drawn from a binary distribution (50% for the choice of 1 of 2 images, 33% in the case of 1 of 3 images, and so forth up to 3% in the case of 1 out of 32 images). This is the most conservative estimate; additional assumptions on the distribution would yield lower p-values.

in Fig. 3 are significantly above their corresponding chance levels, with the least significant point corresponding to predictions using CIO with image set size 4: this results in correct decoding in 33.6% of trials, compared to 25% for chance, with null hypothesis that predictions are 25% correct being rejected at $p < 10^{-4}$.

We also tested the prediction rates when fixations were pooled over progressively fewer subjects, instead of only nine as above. For this, we used only the CIO+F map (although the face channel shows the highest decoding performance we wanted to use a feature map that combines bottom-up features to match common attention prediction methods), and binary image trials (one decoy). One might imagine that pooling over fixation recordings from different subjects

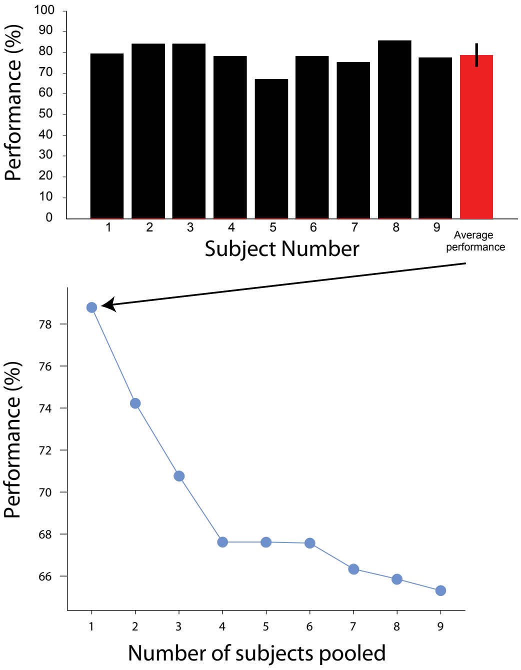

Fig. 4. Performance of the 9 individual subjects. **Upper panel.** For the 441 scanpaths/images, we computed the decoding performance of each individual subject. Bars indicate the performance of each subject. Red bar on the right indicates the average performance of all 9 subjects, with standard error bar. Average subject performance was 79%, with the lowest decoding performance at 67% (subject 4), and the highest at 86% (subject 8). All values are significantly above chance (50%), with p values (sign test) below 10^{-10}. **Lower panel.** Performance of various combinations of the 9 subjects. Scanpaths of 1, 2, ... 9 subjects used to determine the performance differences by using average scanpaths of multiple subjects. The performance of individual subjects shown on the leftmost point is the average of each subjects' performance as shown in the upper panel. The rightmost point is the performance of all subjects combined. Each subject pool was combined from a random choice of subjects out of the 9, reaching the pool size.

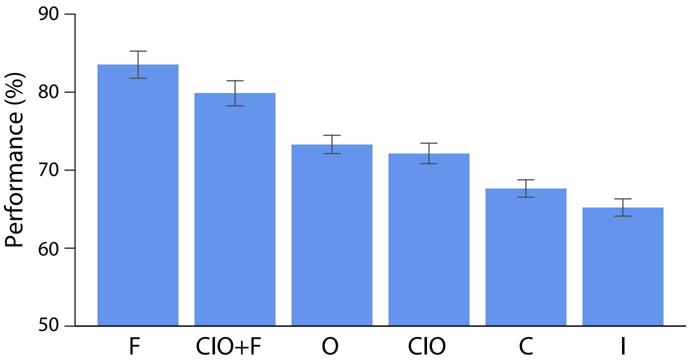

Fig. 5. Decoding performance based on feature maps used. We show the average decoding performance on binary trials using each of the 6 different feature maps, and in each trial, the scanpath of only one individual subject. Thus, for instance, the performance of the CIO+F map is exactly that shown in the average bar in Fig. 4. The higher the performance the more useful the feature is in the decoding. The face channel is the most important one for this dataset.

would increase the signal to noise ratio, but in fact we find that prediction performance only decreases (Fig. 4) with more subjects. There are several possible explanations for this decrease. First, in computing the AUC, we record a correct detection ("hit") whenever a superthreshold saliency map cell overlaps with at least one fixation, but discard information about multiple fixations at that location (*i.e.*, a cell is either occupied by a fixation or not). Thus, the accuracy of the ROC AUC agreement between a saliency map and the fixations of multiple observers degrades with overlapping fixations. As the number of overlapping fixations increases with observers, the reliability of our decoding measure decreases. Indeed, other measures taking into account this phenomenon then can outperform the present metric. Second, if different observers exhibit distinct feature preferences (say, some prefer "color", some prefer "orientation", etc.), the variability in the locations of such features across an image set would contribute to the prediction in this set. It is possible that an image set is more varied along the preferences of any one observer on average than along the pooled preferences of multiple observers. This would make it more difficult to decode from aggregate fixation sets.

The mean percentage of correct decoding for a single subject was 79% (chance is 50%), ($p < 10^{-288}$, sign test). For all combinations of 1 to 9 subjects used, the prediction was above chance (with p values below $p < 10^{-10}$). The lowest prediction performance results from pooling over all nine subjects, with 66% hit rate (still significantly above chance at 50%). Figure 4 shows the prediction for each of the 9 subjects with the CIO+F feature map.

Finally, in order to test the relative contribution of each feature map to the decoding, we used our new decoding correctness rate to compare feature map types, from most discriminating to least. This was done by comparing separately each of the 6 features maps' average decoding performance for binary trials with 9 individual subjects' scanpaths. The results (Fig. 5) show that out of the 6 feature maps the face channel has the highest performance (decoding performance of 82%, $p = 0$) (as shown also in Fig. 3), and the intensity map has the lowest performance (decoding performance: 65%, $p < 10^{-104}$, sign test). All values are significantly above chance (50%).

4 Discussion

In this study, we investigated if scanpath data could be used to decode which image an observer was viewing given only the scanpath and saliency maps. The results were quite strong: in an experiment with 441 trials, each consisting of 32 images with scanpath data belonging to one unknown image in the set, in 73 trials (17%) the correct image was selected, a fraction much higher than chance ($\frac{1}{32} = 3\%$). This leads us to propose a new metric for quantifying the efficacy of saliency maps based on image discriminability. For decoding we used the standard area under ROC curve measure with the fixations from 1 to 9 subjects on a feature map generated by popular models for fixations and attention predictions.

The "decodability" of a dataset is a score given to the combined scanpath/stimuli data for a given feature and as such can be used in various ways: we here used the decodability in order to compare ideal combined subjects' scanpath pool and feature maps' predictive power. Furthermore, we can imagine the same method being used to cluster subjects according to features that pertain specifically to them for a given dataset (*i.e.* if a particular set of subjects tends to look more often on an area in the images than other [19], or tends to fixate on a certain object/target more [20,21,22], this would result in a higher decoding performance for that feature map), or as a measure of the relative amount of stimuli needed to reach a certain level of decoding performance. Our data suggests that clustering by such features to segregate between autistic and normal subjects is perhaps possible based on differences in their looking at faces/objects [21]. However, our autism subjects fixations dataset is too small to reach significance.

In line with earlier results, ours show that saliency maps using bottom-up features such as color, orientation, and intensity are relatively accurate predictors of fixation [16,23,24,25,26] with a performance above 70% (Fig. 5, similar to the estimate in [15]). Adding the information from a face detector boosts performance to over 80%, similar to the estimate in [17]. It is possible that incorporating more complex, higher-level feature maps [27,28] could further improve performance.

Some of the images we used were very similar to each other, and so the image set could be considered challenging. Using this novel decoding metric on larger, more diverse datasets could yield more striking distinctions between the feature maps and their relative contributions to attentional allocation.

Notice that in the results, in particular in Fig. 3, we computed average predictive performance using fixations pooled over all 9 scanpaths recorded per image. However, as we have shown that individual subjects' fixations are more predictive due to variability issues, these results should be even stronger than those we have included above.

A possibility for subsequent work is the prediction not of particular images from a set, but of image content. For example, is it possible to predict whether or not an image contains a face, text, or other specific semantic content based only on the scanpaths of subjects? The same kinds of stereotypical patterns we used to predict images would be useful in this kind of experiment.

Finally, one can think of more sophisticated algorithms for predicting scanpath/image pairs. For instance, one could use information about previously decoded images for future iterations (perhaps by eliminating already decoded images from the pool, making harder decoding more feasible), or a softer ranking algorithm (here we considered decoding correct only if the corresponding scanpath was ranked the highest among 32 images; one could, however, compute statistics from a soft "confusion matrix" containing all rankings so as to reduce the noise from spuriously high similarity pairs).

We demonstrated a novel method for estimating the similarity between a given set of scanpaths and images by measuring how well scanpaths could decode the images that corresponded to them. Our decoder ranked images according to saliency map/fixation similarity, yielding the most similar image as its prediction. While our decoder already yields high performance, there are more sophisticated distance measures that might be more accurate, such as ones used in electrophysiology [7].

Rating a saliency map relative to a scanpath based on its usability as a decoder for the input stimulus represents a robust new measure of saliency map efficacy, as it incorporates information about how dissimilar a map is from those computed on other images. This novel method can also be used for assessing images sets, for measuring the performance and attention allocation for a given set, for comparing existing saliency map performance measures, and as a metric for the evaluation of eye-tracking data against other psychophysical data.

Acknowledgements

This research was funded by the Mathers Foundation, NGA and NIMH.

References

1. Young, M., Yamane, S.: Sparse population coding of faces in the inferotemporal cortex. Science 256(5061), 1327–1331 (1992)
2. Schwartz, E., Desimone, R., Albright, T., Gross, C.: Shape Recognition and Inferior Temporal Neurons. Proceedings of the National Academy of Sciences of the United States of America 80(18), 5776–5778 (1983)
3. Sato, T., Kawamura, T., Iwai, E.: Responsiveness of inferotemporal single units to visual pattern stimuli in monkeys performing discrimination. Experimental Brain Research 38(3), 313–319 (1980)
4. Perrett, D., Rolls, E., Caan, W.: Visual neurones responsive to faces in the monkey temporal cortex. Experimental Brain Research 47(3), 329–342 (1982)

5. Logothetis, N., Pauls, J., Poggio, T.: Shape representation in the inferior temporal cortex of monkeys. Current Biology 5(5), 552–563 (1995)
6. Hung, C., Kreiman, G., Poggio, T., DiCarlo, J.: Fast Readout of Object Identity from Macaque Inferior Temporal Cortex (2005)
7. Quiroga, R., Reddy, L., Koch, C., Fried, I.: Decoding Visual Inputs From Multiple Neurons in the Human Temporal Lobe. Journal of Neurophysiology 98(4), 1997 (2007)
8. Itti, L., Koch, C., Niebur, E., et al.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
9. Dickinson, S., Christensen, H., Tsotsos, J., Olofsson, G.: Active object recognition integrating attention and viewpoint control. Computer Vision and Image Understanding 67(3), 239–260 (1997)
10. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. 4(4), 219–227 (1985)
11. Yarbus, A.: Eye Movements and Vision. Plenum Press, New York (1967)
12. Goldstein, R., Woods, R., Peli, E.: Where people look when watching movies: Do all viewers look at the same place? Computers in Biology and Medicine 37(7), 957–964 (2007)
13. Privitera, C., Stark, L.: Evaluating image processing algorithms that predict regions of interest. Pattern Recognition Letters 19(11), 1037–1043 (1998)
14. Itti, L., Koch, C.: Computational modeling of visual attention. Nature Rev. Neurosci. 2(3), 194–203 (2001)
15. Peters, R., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision Res. 45(18), 2397–2416 (2005)
16. Tatler, B., Baddeley, R., Gilchrist, I.: Visual correlates of fixation selection: effects of scale and time. Vision Research 45(5), 643–659 (2005)
17. Cerf, M., Harel, J., Einhäuser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems, vol. 20. MIT Press, Cambridge (2008)
18. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. Computer Vision and Pattern Recognition 1, 511–518 (2001)
19. Buswell, G.: How People Look at Pictures: A Study of the Psychology of Perception in Art. The University of Chicago press (1935)
20. Barton, J.: Disorders of face perception and recognition. Neurol. Clin. 21(2), 521–548 (2003)
21. Klin, A., Jones, W., Schultz, R., Volkmar, F., Cohen, D.: Visual Fixation Patterns During Viewing of Naturalistic Social Situations as Predictors of Social Competence in Individuals With Autism (2002)
22. Adolphs, R.: Neural systems for recognizing emotion. Curr. Op. Neurobiol. 12(2), 169–177 (2002)
23. Baddeley, R., Tatler, B.: High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. Vision Research 46(18), 2824–2833 (2006)
24. Einhäuser, W., König, P.: Does luminance-contrast contribute to a saliency map for overt visual attention?. Eur. J. Neurosci. 17(5), 1089–1097 (2003)
25. Einhäuser, W., Kruse, W., Hoffmann, K., König, P.: Differences of monkey and human overt attention under natural conditions. Vision Res. 46(8-9), 1194–1209 (2006)

26. Navalpakkam, V., Itti, L.: Search goal tunes visual features optimally. Neuron 53(4), 605–617 (2007)
27. Kayser, C., Nielsen, K., Logothetis, N.: Fixations in natural scenes: Interaction of image structure and image content. Vision Res. 46(16), 2535–2545 (2006)
28. Einhäuser, W., Rutishauser, U., Frady, E., Nadler, S., König, P., Koch, C.: The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. J. Vis. 6(11), 1148–1158 (2006)