Predicting observers' task from their scanpaths on natural scenes

Ali Borji and Laurent Itti {borji,itti}@usc.edu
University of Southern California, Computer Science, Psychology, and Neuroscience

http://ilab.usc.edu/publications/doc/Borji_Itti14vss.pdf

Supported by the National Science Foundation,
the Army Research Office, and the U.S. Army

(1) Introduction & Motivation

While the hypothesis that it is possible to decode the observer's task from eye movements (Yarbus 1967) has received some support (e.g., DeAngelus & Pelz, 2009; Henderson, Shinkareva, Wang, Luke, & Olejarczyk, 2013; Iqbal & Bailey, 2004), Greene, Liu, and Wolfe (2012) argued against it by reporting a failure.

In this study, we perform a more systematic investigation of this problem, probing a larger number of experimental factors than previously. Our main goal is to determine the informativeness of eye movements for task and mental state decoding.

(2) Related works

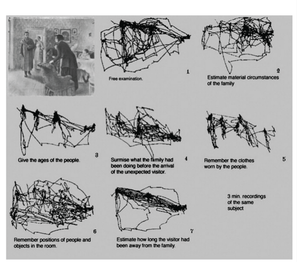
Example stimuli for the A) scene memorization task, B) reading task, C) scene search task and D) pseudo-reading task.

Decoding Accuracy > 64 % (chance = 25 %) using multivariate pattern classification.

The features were the mean and standard deviation of fixation duration, the mean and standard deviation of saccade amplitude, the number of fixations per trial, and the three parameters μ, σ, and τ quantifying the shape of the fixation duration distribution with an ex-Gaussian distribution.

Henderson et al., PLOS ONE 2013

TASKS:
1. Memorize the picture (memory).
2. Determine the decade in which the picture was taken (decade).
3. Determine how well the people in the picture know each other (people).
4. Determine the wealth of the people in the picture (wealth).



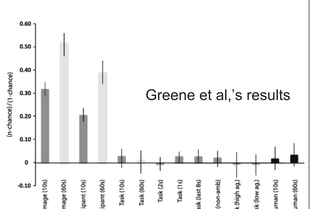
Yarbus' experiment on Repin's painting

Decoding Accuracy:
- **viewed image:** 33% correct vs. chance = 1.5% (64 images)

- **participant:** 26% vs. chance = 6.3% (16 participants)

- **task:** 27.1% correct vs. chance = 25% (4 tasks) ?! n.s.

The features were (1) number of fixations, (2) the mean fixation duration, (3) mean saccade amplitude, (4) percent of image covered by fixations assuming a 1° fovea, dwell-time on various regions of interest: (5) faces, (6) human bodies, and (7) objects.



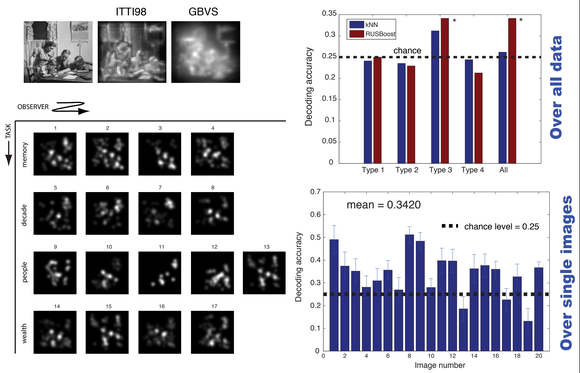
Greene et al,'s results

Greene et al., Vis. Res. 2012

(3) Experiment 1

Classifiers:
- k-nearest-neighbor; kNN
- RUSBoost (random undersampling boost) algorithm (Seiffert, Khoshgoftaar, Van Hülse, & Napolitano, 2010)

Features:
Type 1: the smoothed fixation map, down sampled to 100 × 100 (10000 D)
Type 2: histograms of normalized scan path saliency (NSS) (9 x 70 = 630 D)
Type 3: first 4 features of Greene et al. (4D)
Type 4: < x, y > locations of the first five fixations (i.e., a 10 D vector)



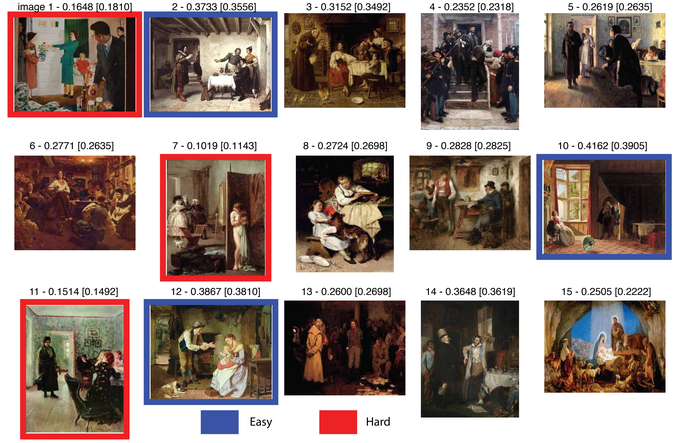
Decoding Accuracy: [over all data with Bonferroni correction]
Using kNN and Feature Type 3, we achieved accuracy of 0.3118, which is above Greene et al.'s (2012) results and is significantly better than chance (k = 8; p = 0.014).

Using RUSBoost and Feature Type 3, we achieved accuracy of 0.3412 (p = 1.07e – 04).

(4) Experiment 2

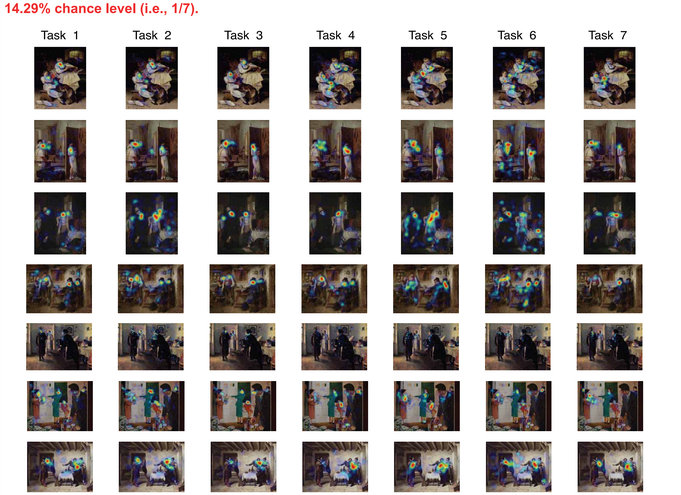
Observers: 21 USC students (10 male, 11 female)
Visual angles: 43 x 25 degrees
Stimuli: 15 paintings (13 oil on canvas, some by Repin)
Viewing time: 30 seconds

Images 1–5	6–10	11–15
3 0 × T 1	3 0 × T 2	3 0 × T 3
3 0 × T 2	3 0 × T 3	3 0 × T 4
3 0 × T 3	3 0 × T 4	3 0 × T 5
3 0 × T 4	3 0 × T 5	…
3 0 × T 5	…	…
3 0 × T 6	…	…
3 0 × T 7	3 0 × T 1	3 0 × T 2



Decoding Accuracy: [over all data with Bonferroni correction]
We report results using a leave-one-out procedure. We have 21 observers, each viewing 15 images (each five images under a different question; three questions per observer) thus resulting in 315 scan paths.
Using Feature Type 1, we achieved average accuracy of 0.2421, which is significantly above chance (binomial test, p = 2.4535e – 06). Using Feature Type 2 (i.e., NSS histogram of nine saliency models as in Experiment 1) results in accuracy of 0.2254 (p = 5.6044e – 05).

Decoding Accuracy: [over single images with Bonferroni correction]
Three observers viewed each image under one question thus resulting in 21 data points per image (i.e., 3 Observers × 7 Questions). Note that each set of three observers were assigned the same question.
RUSBoost classifier and Feature Type 1 results in average accuracy of 0.2724 over 50 runs and 15 images. Using first two feature types (a 10,000 + 23 × 70 = 11610D vector) results in average performance of 0.2743. Over all runs (i.e., table rows), the minimum accuracy (average over all 15 images) is 0.2540 and maximum accuracy is 0.3079. Note that our accuracies are almost two times higher than the 14.29% chance level (i.e., 1/7).

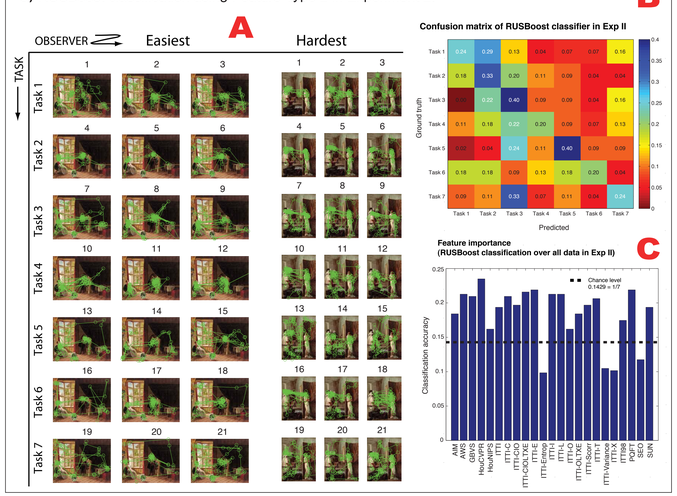


(5) Experiment 2 (cntd)

A) Confusion matrix averaged over 15 images and 50 RUSBoost runs using Feature Type 1. We observe high accuracies for Task 3 (estimating age), Task 5 (remembering cloths), and Task 7 (estimating how long the visitor has been away) but low accuracy for the free-viewing task. There is a high confusion between Task 2 and Tasks 6 and 1 and also between Task 1 and Task 7.

B) The easiest and hardest stimuli using Feature Type 1.

C) RUSBoost classification using Feature Type 2 in Experiment 2.



Conclusions & Discussions

1) Successful task decoding results provide further evidence that fixations convey diagnostic information regarding the observer's mental state and task, consistent with the cognitive relevance theory of attention (see Hayhoe & Ballard, 2005).

2) It is possible to reliably infer the observer's task from Greene et al.'s (2012) data using stronger classifiers. Classification was better when we treated images individually.

3) Is it always possible to decode task from eye movements? We argue that there is no general answer to this type of pattern recognition questions. Answers depend on the used stimuli, observers, and questions.

4) Here we followed the procedure by Greene et al. (2012) in which: (a) no observer viewed the same image twice and (b) the same scene was shown under multiple questions. The first rule aims to eliminate memory biases. The second rule ensures that the final result is not due to differences in stimuli.

5) Recently Kanan et al. (2014), were also able to decode the task from eye movement patterns on Greene et al.,'s data with about the same performance as us (> 33 %).

6) Beyond scientific value, decoding task from eye movements has practical applications such as wearable visual technologies (e.g., Google glass) and patient diagnosis (e.g., ASD, ADHD, FASD, Parkinson's, and Alzheimer's).

[1] DeAngelus M., Pelz J. B. (2009). Top-down control of eye movements: Yarbus revisited. Visual Cognition, 17, 790–811.
[2] Kanan C., Ray N., Bseiso D. N. F., Hsiao J. H., Cottrell G. W. (2014). Predicting an observer's task using multi-fixation pattern analysis. In Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA-2014).
[3] Yarbus A. L. (1967). Eye movements and vision. New York: Plenum.
[4] Iqbal S., Bailey B. (2004). Using eye gaze patterns to identify user tasks. Proceedings of the Grace Hopper Celebration of Women in Computing, October 6–9, 2004.
[5] Henderson J., Shinkareva S., Wang J., Luke S., Olejarczyk J. (2013). Predicting cognitive state from eye movements. Plos One, 8 (5), e64937.
[6] Itti L., Koch C., Niebur E, (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 1254–1259.
[7] Seiffert C., Khoshgoftaar T. M., Van Hülse J., Napolitano A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 40, 185–197.
[8] Borji, A., and L. Itti. Defending Yarbus: Eye movements reveal observers' task. Journal of vision 14,3 (2014): 29.
[9] Hayhoe M., Ballard D. (2005). Eye movements in natural behavior. Trends in Cognitive Sciences, 9, 188–194.