Faces and text attract gaze independent of the task: Experimental data and computer model

Moran Cerf
Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA

E. Paxon Frady
Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA

Christof Koch
Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA, & Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

Previous studies of eye gaze have shown that when looking at images containing human faces, observers tend to rapidly focus on the facial regions. But is this true of other high-level image features as well? We here investigate the extent to which natural scenes containing faces, text elements, and cell phones—as a suitable control—attract attention by tracking the eye movements of subjects in two types of tasks—free viewing and search. We observed that subjects in free-viewing conditions look at faces and text 16.6 and 11.1 times more than similar regions normalized for size and position of the face and text. In terms of attracting gaze, text is almost as effective as faces. Furthermore, it is difficult to avoid looking at faces and text even when doing so imposes a cost. We also found that subjects took longer in making their initial saccade when they were told to avoid faces/text and their saccades landed on a non-face/non-text object. We refine a well-known bottom–up computer model of saliency-driven attention that includes conspicuity maps for color, orientation, and intensity by adding high-level semantic information (i.e., the location of faces or text) and demonstrate that this significantly improves the ability to predict eye fixations in natural images. Our enhanced model's predictions yield an area under the ROC curve over 84% for images that contain faces or text when compared against the actual fixation pattern of subjects. This suggests that the primate visual system allocates attention using such an enhanced saliency map.

Keywords: attention, eye tracking, faces, text, natural scenes, saliency model, human psychophysics

Citation: Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*(12):10, 1–15, http://journalofvision.org/9/12/10/, doi:10.1167/9.12.10.

Introduction

Visual attention serves to delegate the resources of the brain to quickly and efficiently process the vast amount of information that is available in the environment (James, 1890). Certain aspects of selective visual attention, in particular task-independent, exogenous and bottom–up driven attention, has become reasonably well understood, and quantitative models have been derived to explain attentional and eye movement deployments in the visual scene (Itti & Koch, 2001). In general, it is thought that observers' fixational eye patterns correlate tightly with their covert attention under natural viewing conditions (Einhäuser, Kruse, Hoffmann, & König, 2006; Parkhurst, Law, & Niebur, 2002; Rizzolatti, Riggio, Dascola, & Umilta, 1987).

Commonalities between different individuals' fixation patterns allow computational models to predict where people look and the order in which they view different items (Cerf, Cleary, Peters, Einhäuser, & Koch, 2007; Foulsham & Underwood, 2008; Oliva, Torralba, Castelhano, & Henderson, 2003). There are several models for predicting observers' fixations inspired by putative neural mechanisms (Dickinson, Christensen, Tsotsos, & Olofsson, 1994). One of the dominant sensory-driven models of attention focuses on low-level attributes of the visual scene to evaluate the most salient areas. Features such as intensity, orientation, and color are combined to produce maps through center-surround filtering at multi-scaled resolutions. These maps of feature contrast are normalized and combined to create an overall saliency map, which predicts human fixations significantly above chance (Itti & Koch, 2000). Filling some of the gaps between the predictive power of current saliency map models and their theoretical optimum is the incorporation of higher order statistics (Einhäuser, Rutishauser, et al., 2006). One way of doing this is by adding new feature channels for faces or text into the saliency map.

Visual attention is frequently deployed to faces, to the detriment of other visual stimuli (Bindemann, Burton, Hooge, Jenkins, & de Haan, 2005; Bindemann, Burton, Langton,

doi: 10.1167/9.12.10 Received November 10, 2008; published November 18, 2009 ISSN 1534-7362

Schweinberger, & Doherty, 2007; Cerf, Harel, Einhauser, & Koch, 2008; Mack, Pappas, Silverman, & Gay, 2002; Ro, Russell, & Lavie, 2001; Theeuwes & Van der Stigchel, 2006; Vuilleumier, 2000). Evidence from infants as young as 6 weeks old suggests that faces are visually captivating (Cashon & Cohen, 2003). We here investigate the attractiveness of faces in the context of detecting a target in natural scenes where faces are embedded in the images.

Text is yet another entity that frequently captures humans' gaze in natural scenes (Cerf, Frady, & Koch, 2008). Although the claim that one is born with an innate ability to detect faces is still under debate (Golarai et al., 2007; Simion & Shimojo, 2006), it is unlikely that we are born with a similar mechanism for text or cell phones detection. Any skill in detecting these man-made items in a natural scene should be attributable to experience.

This study considers how faces, text, and complex man-made objects (cell phones) affect gaze in still images. We track subjects' eye movements in a free-viewing task and in multiple search tasks to measure the extent to which subjects can avoid looking at faces and other objects. Such tasks are close to day-to-day situations and shed light on the way attention is allocated to important semantic entities.

Materials and methods

Experimental procedures

Subjects viewed a set of images (1024 × 768 pixels) in four experiments. The general structure of all tasks is the same. Prior to each session, the subjects' gaze was determined through a calibration process. Before each stimulus onset, the subjects were instructed to look at a white cross at the center of a gray screen. If the calculated gaze position was not at the center of the screen, the calibration process was repeated to ensure that position was consistent throughout the experiment. Eye-position data were acquired at 1000 Hz using an Eyelink 1000 (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 (Brainard, 1997; Cornelissen, Peters, & Palmer, 2002). 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° × 21°. Subjects used a chin rest to stabilize their head. Eye movement data were acquired from the right eye alone. All subjects had normal or corrected-to-normal eyesight. All subjects were naïve to the purpose of the experiment. These experiments were undertaken with the understanding and written consent of each subject. All experimental procedures were approved by Caltech's Institutional Review Board.

In the first ("free-viewing") experiment, 27 subjects viewed one out of three categories of images, with nine subjects per category. The categories were natural scenes that contained one or more faces (157 images), one or more discrete text elements (37 images), or one cell phone (37 images). Images were presented to the subjects for 2 seconds, after which they were instructed to answer the question, "How interesting was the previous image?" using 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. Image order within each block was randomized throughout the experiment. Figure 1a shows a sample image from each of the three categories.

In order to provide a fully equivalent image group, for comparison between the categories independent of size and background, we had six additional subjects perform Experiment 2 ("control for the relative effect of size"), where a set of 25 images was presented in a "free-viewing" task. These images had faces, text, or cell phones artificially embedded such that they occupied the exact same size and location in the image.

In a third ("search") experiment, 15 additional subjects were instructed to look for a target—two concentric circles (50 × 50 pixels; 1.36° × 1.36° visual angle) surrounding a cross-hair (Figure 1b)—embedded in the image. The cross's location was selected randomly from a uniform distribution. Each block had 74 images from a single category, 37 of the images were the same as in the "free-viewing" experiment, and the other 37 were identical images except that they included the cross embedded somewhere within the image. Images were presented for 2 seconds after which subjects were given a two-alternative-forced-choice (2AFC) question asking "Was the cross in the previous image (y/n)." Seven of the 15 subjects performed under a "free search" instruction, in which they were told that "The target can be located anywhere in the following images." The remaining eight subjects performed the search task under an "avoid" instruction, in which they were told, "The target cannot be located on face objects in the following images" for the face block, "text objects" for the text block, and "cell phone objects" for the cell phone block. The target was placed outside of these regions accordingly. Before the experiment, subjects were given 12 training trials using separate images. Subjects were told the location of the cross at the end of each practice trial in order to familiarize them with the looks of the cross-hair and difficulty level.

Finally, four additional subjects performed a search task in a 4th experiment ("control for the effect of adaptation"), where images from each category were intermixed, creating one long experiment. The experiments included both the free search task and avoid task instructions such that prior to each trial the instructions could be either "The target will not appear on a face," "on a text," or "on a cell phone object" or "The target can appear anywhere."

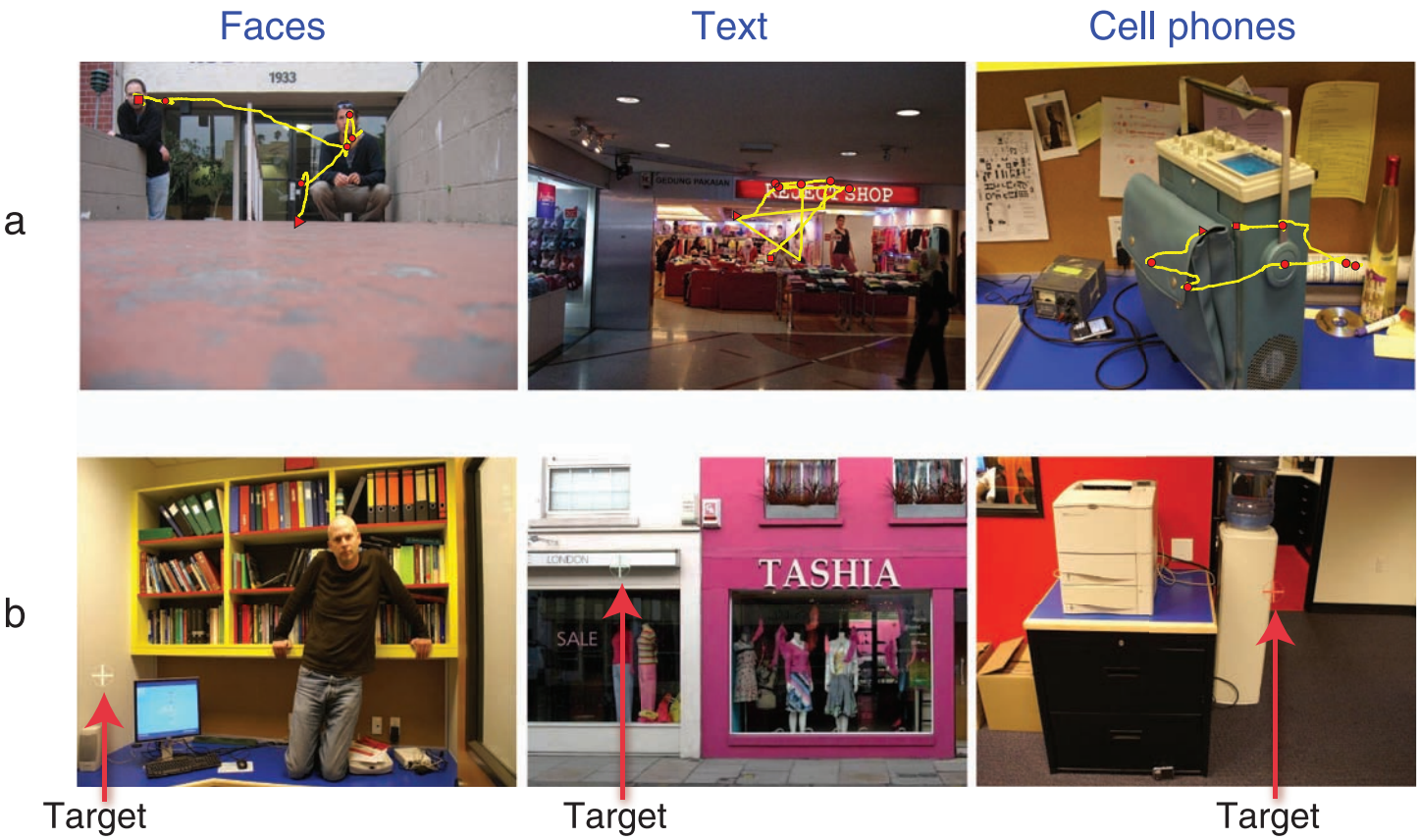


Figure 1. (a) Examples of images from the three categories: faces, text, and cell phones, with scanpaths of one individual from each of the three groups superimposed. The triangle marks the first and the square the last fixation, the yellow line the scanpath, and the red circles the subsequent fixations. The trend of visiting the faces and text first—typically within the 1st or 2nd fixation—is evident, while cell phones do not draw eye gaze in the same manner. (b) Examples of images used in Experiment 3 ("search"). Red arrows point to the superimposed target cross.

This experiment controlled for a general adaptation of strategy that may occur over a block of the same image types. Furthermore, by looking only at the free search trials in this experiment, we can rule out any possible top–down influences as the subjects are unaware of the coming stimulus category prior to the image presentation.

Images

All stimuli were designed or chosen as images that are representative of a real-world scene.

The face images were manually photographed in indoor and outdoor environments. Images included people (and their faces) of various skin colors, age, and postural positions. A few images had face-like objects (smiley T-shirt, animal faces, and objects that had irregular faces in them—masks or the Egyptian Sphinx). The text images were taken from the Internet and were chosen such that only a few text objects appeared in the scene. Images containing a cell phone were manually photographed in an office setting where a cell phone would be considered a reasonable item. Cell phones were chosen to represent an object that is less important to a human's visual environment. The majority of the images contained only one of the three entities.

The average face size was 5% ± 4% (5.42° ± 4.84° visual angle); the average size of text was 6% ± 1% (5.93° ± 2.42° visual angle); and the average size of cell phones was 1% ± 0.4% of the entire image (2.42° ± 1.53° visual angle).

In the search tasks, the cross's color was varied such that it would be challenging for the viewer to find, but obvious to recognize once it was located. The color was altered throughout the experiment such that observers could not solve the task by simply looking for a specific color. After choosing a random location for the cross (given the constraints of each block), the cross's color was determined by taking the average of all pixel colors in a local (50 × 50 pixels) neighborhood and then increasing the brightness by 10% (this made the cross visible in all locations in which it was placed, while still making it challenging to find). These parameters were chosen to yield about 70% success rate on 2AFC test trials.

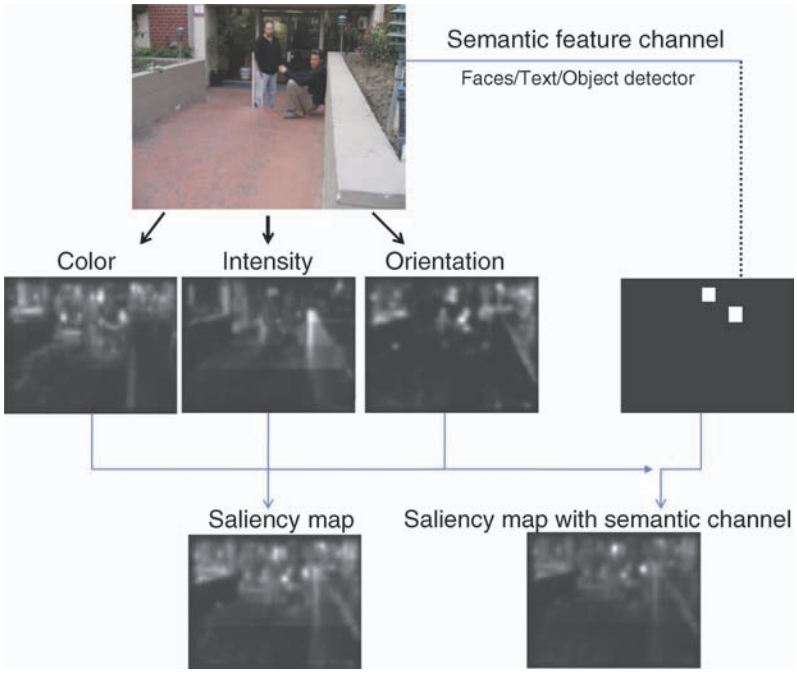


Figure 2. Illustration of the combined saliency map with a semantic channel added. An example image is fed into feature channels for color, intensity, and orientation as well as into a fourth channel for faces/text/cell phone. The combined feature channels were normalized and formed a modified saliency map that was compared to the original saliency map.

All the fixations, scanpaths, and images used in the experiment are available online at http://www.fifadb.com for further studies.

Model

The scanpaths of the subjects in the free-viewing task were used to validate the predictions of subjects' attention allocation by a computer model.

Prior work shows that a biologically inspired bottom–up driven saliency model can predict subjects' fixation to an accuracy of about 53% for images containing natural scenes (Peters, Iyer, Itti, & Koch, 2005). More so, prior work by Cerf, Harel, Einhauser, et al. (2008) showed that inclusion of a face map that is combined with the bottom–up feature maps yields better performance in predicting subjects' fixations.

We here tested the ability to improve fixation prediction by adding a channel containing the exact location of the faces, text, or cell phones in the image. We manually defined a minimally sized "region of interest" (ROI) around each face, text, or cell phone, creating a binary heat-map describing the location of the entities. The original saliency map is computed as an average of three channels: intensity, orientation, and color (Itti, Koch, & Niebur, 1998):

$$S = \frac{1}{3}[N(I) + N(C) + N(O)]. \quad (1)$$

The modified saliency map adds the extra entity channel (E) (for illustration, see Figure 2):

$$S = \frac{1}{4}[N(I) + N(C) + N(O) + N(E)]. \quad (2)$$

The combination was linear with uniform weight for simplicity. Performance of the saliency maps was measured by the receiver operating characteristic (ROC) curve. For each saliency map, the hit rate was calculated by determining the locations where the saliency map was above threshold and a fixation was present in these suprathreshold regions. Similarly, the false alarm rate was calculated by measuring the locations at which the saliency map was above threshold and there was no fixation present (Cerf, Harel, Huth, Einhauser, & Koch, 2008). The ROC curve was generated by varying the threshold to cover all possible ranges of values the saliency map produces. The area under the ROC curve (AUC) is a general measure of how well the saliency map predicts fixations.

The AUCs were normalized by an "ideal AUC," which measures how well the subject's fixations predicted each other. This ideal AUC reflects an upper bound to how well our model can predict subjects' fixations. The AUC normalization was done such that a value of 100% would reflect the ideal AUC and a value of 50% would reflect chance. The ideal AUC was calculated by performing a similar ROC analysis on each subject using the fixation pattern of all other subjects in place of the saliency map.

This leave-one-out analysis results in an ideal AUC of 78.6% ± 6.1% for faces, 78.4% ± 5.5% for text, and 79.2% ± 6.0% for cell phones under the free-viewing condition. None of these values are significantly different from another. This shows that the inter-subject variability is consistent for the image sets, suggesting that there is little difference between the visual complexities of the image sets.

Analysis metrics

Fixations

Fixations were determined by the built-in software of our eye-tracking system. We categorized fixations by their "fixation number" based on a fixation's position in the ordered sequence of fixations (i.e., first, second, third). The "initial fixation" is the fixation occurring before stimulus onset, when the subjects are focusing on the centered fixation cross, and is not counted as part of the ordered sequence of fixations.

We calculate the fraction of fixations in an ROI by summing the number of fixations that fall within this region over all trials and then dividing by the number of trials. A trial consists of one subject being presented one image and will contain one ordered sequence of fixations.

Saccades

Saccades were also determined by the eye-tracking system. The "saccade planning time" is the duration of time between the stimulus onset and the initiation of the first saccade. Saccade planning times smaller than 50 ms or greater than 600 ms were discarded to remove outliers and artifacts. The duration of viewing time was measured based on when the saccade started as opposed to when the next fixation started so that timing would not be affected by the length of the saccade or the distance to the target.

Saccades were categorized as being "on-target" or "off-target" depending on the location of the following fixation. If the fixation fell in an ROI (i.e., on a face), then the saccade was considered to be on-target. If the fixation fell outside a region of interest, then the saccade was categorized as off-target (for illustration, see Figure 3).

Baseline calculation

To compute chance level of performance, we calculated the ratio of the fraction of fixations that land in a target's region of interest to the fraction of a baseline distribution. The baseline for a particular image is the fraction of all subjects' fixations from all other images that fall in the

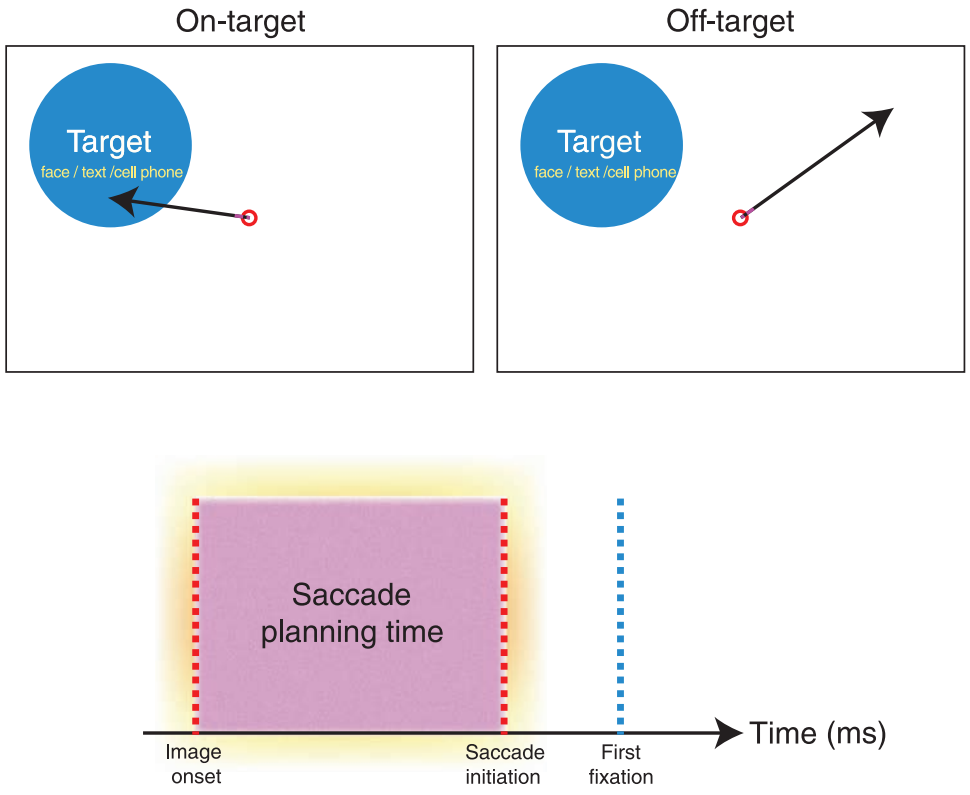


Figure 3. Illustration of the computation of the saccade planning time. For each subject and image, we take the time spent before the initiation of the first fixation as the "saccade planning time." The location of the saccade's ending point is used to bin the times into two groups based on whether or not the saccade landed in a region of interest. If the first fixation was on a face / text / cell phone we define it as on-target and if it is anywhere else we define it as off-target.

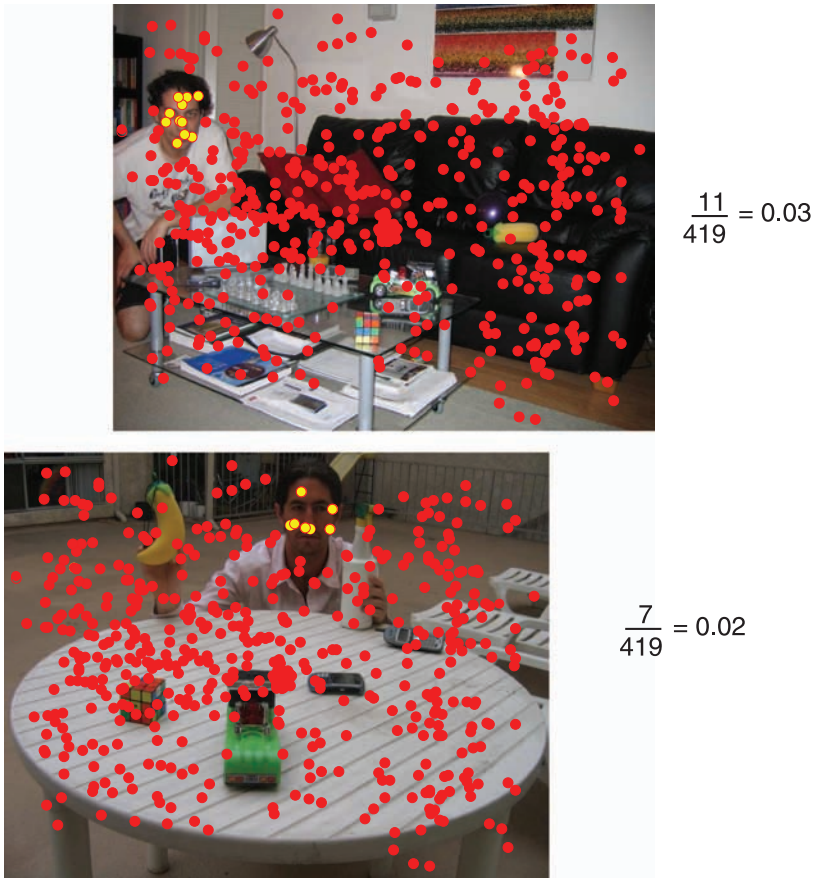


Figure 4. Illustration of the computation of the baseline. For each subject, we consider all fixations, except the ones recorded for this image. We then compute the fraction of these that fall within the ROI for this image. Here we see the baseline calculation for two images containing faces. For each image, the fixations from all other images for that subject are superimposed in red and yellow. Yellow dots indicate that the fixation falls within the ROI. For the bottom image, seven fixations out of 419 fell within the ROI. The average number of fixations in the ROI for all face images using this baseline calculation was 0.4%.

ROI of the particular image (for illustration of the baseline calculation, see Figure 4). This takes into account the varying size and locations of the ROI in all images (as these factors both influence how likely a certain region is to be fixated on by chance) and the double spatial bias of photographer and observer (Tatler, Baddeley, & Gilchrist, 2005). As the baseline value estimates the chance level of landing in a given ROI, we make a comparison with this baseline value to the data by dividing the fraction of fixations landing in the ROI by its baseline value.

Results

Psychophysical results

Experiment 1 ("free viewing")

To evaluate the results of the 27 subjects' free viewing of the images, we looked at both the fixations and the saccade planning times of each subject. The locations of the fixations were compared with minimally sized rectangular ROIs manually defined around each target object—face, text, and cell phone in each image in the entire collection.

During free viewing, subjects fixated on faces within the first two fixations in 89.3% of the trials, which is significantly above chance ($p < 10^{-15}$, *t*-test; Figure 5). Similar measures for the text show that subjects fixated on text within the first two fixations in 65.1% of the trials ($p < 10^{-15}$, *t*-test). Overall, faces, and text are much more likely to be fixated within the first two fixations than is predicted by chance ($p < 10^{-10}$, *t*-test). In contrast, by the 2nd fixation, cell phones were visited in only 8.4% of the trials. The on-target saccade planning time for face images was very rapid (203 ± 57 ms). Off-target saccade planning time was equally rapid (199 ± 72 ms). On-target saccade planning time for text was not as rapid as for faces but significantly faster than for cell phones (239 ± 54 ms for text; 313 ± 47 ms for cell phones; $p < 10^{-5}$ for faces and text compared to cell phones and faces compared to text,

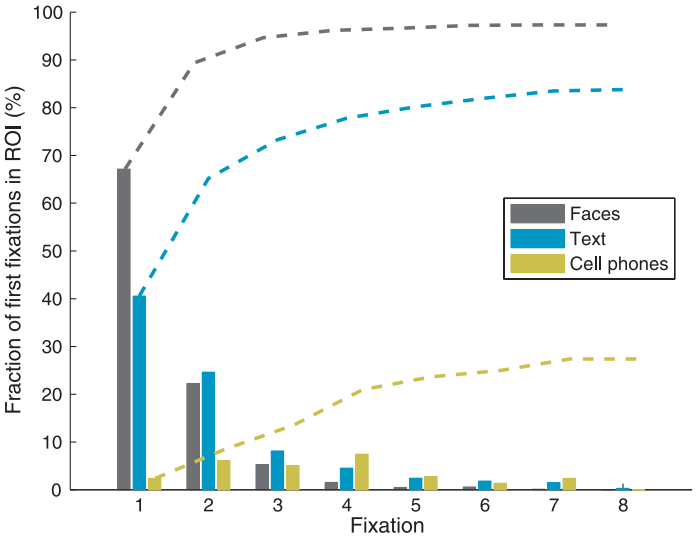


Figure 5. Extent of first fixation on ROI during free-viewing task. Bars depict percentage of trials that reach the ROI the first time in the first, second, third, etc., fixation. The dashed curves depict the integral, i.e., the fraction of trials in which faces were fixated on at least once up to and including the *n*th fixation. These data show that subjects tend to fixate on faces much earlier than text and that they fixate on both text and faces earlier than cell phones.

t-test). Off-target saccade planning time was equally rapid for text images (241 ± 77 ms). Off-target saccade planning time for cell phones was 290 ± 89 ms. Given that the ROIs were chosen very conservatively (i.e., fixations just next to a face did not count as fixations on the face), these results show that faces and text are highly attractive.

Experiment 2 ("control for the relative effect of size")

To be certain that the attractiveness of faces, text, and cell phones is not due to size, position, or competing stimuli in the background, we ran a second, follow-up experiment controlling for each of these factors. Six new subjects were shown a data set comprised of the three entities artificially embedded in the same images (background-wise), such that the entities occupied the same size and position. Subjects engaging in free viewing of these images repeat the same tendency to look at faces the most (61% of the first fixation went to faces), more than text (44% of first fixation landed on a text region), and much more than cell phones (7%). Comparing these results to chance performance, calculated using our baseline distribution of fixations, we see that faces are 16.6 times more likely to be fixated on than the baseline, while text is 11.1 times more likely, and cell phones are only 3.2. There is a significant difference between faces and text ($p < 10^{-6}$, *t*-test), faces and cell phones ($p < 10^{-10}$, *t*-test), and text and cell phones ($p < 10^{-8}$, *t*-test). This strengthens the claim that faces are indeed more attractive to viewers than text, which in turn is more attractive than cell phones. The large difference between the three entities shows that the bias toward faces is due to the higher attractiveness of these, as the shared background among all entities makes the comparison controlled.

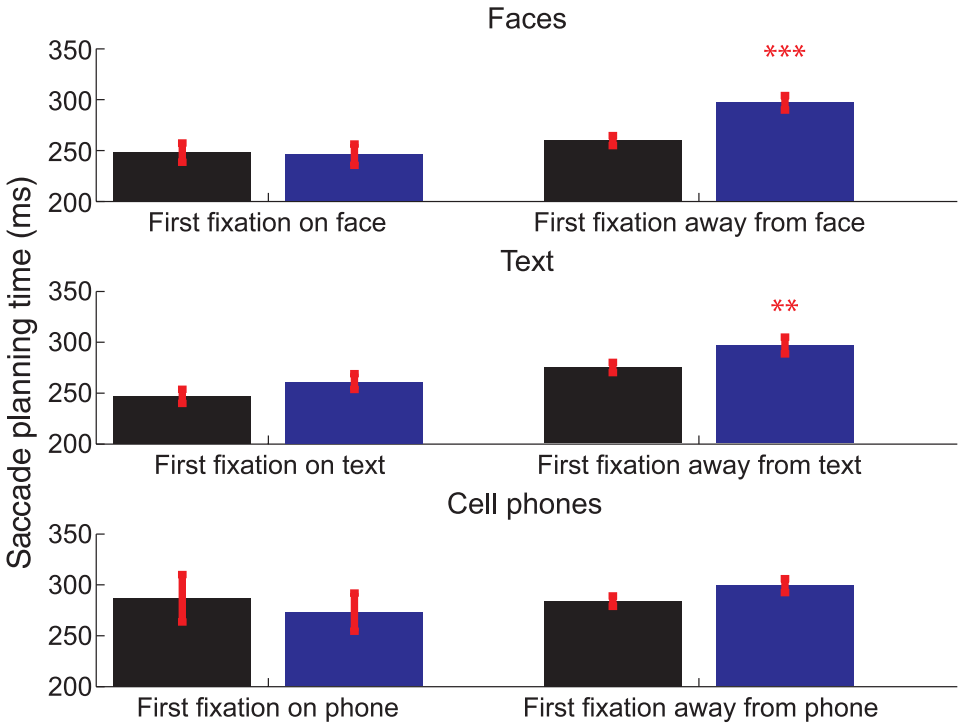
We observe similar trends in the saccade planning times for this experiment for faces (220 ± 51 ms for on-target; 208 ± 66 ms for off-target), for text (233 ± 59 ms, on-target; 236 ± 65 ms, off-target), and for cell phones (299 ± 53 ms, on-target; 310 ± 42 ms, off-target). Again we see that there is a global depression in saccade planning time when there is a face in the image, regardless of whether the saccade is on- or off-target.

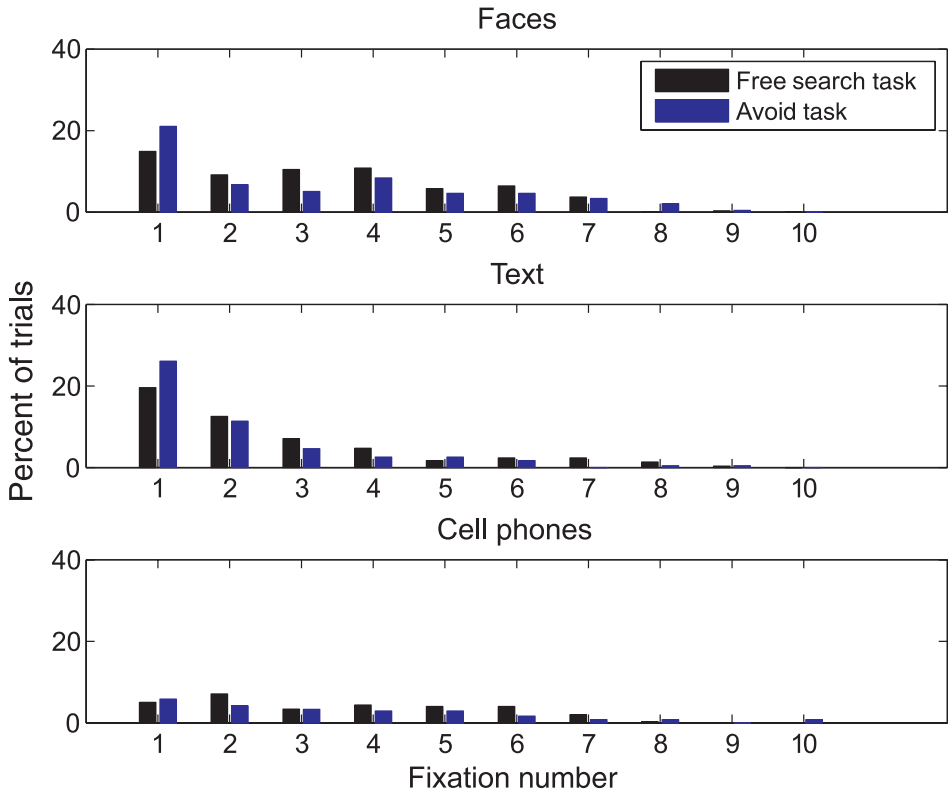
Experiment 3 ("search")

To assess what affect the two different search tasks ("free search" and "avoid") had on subjects' allocation of gaze, we compared fixational patterns between the two tasks. In the free search task, faces were fixated within the first two fixations 24.0% of the time. When subjects were instructed to avoid looking at a face, faces were nevertheless fixated upon within the first two fixations 27.7% of the time. Text was fixated within the first two fixations 32.1% of the time in the free search task versus 37.4% of the time in the avoid task. Cell phones were fixated at approximately the same frequency for both search tasks-12.2% for free search and 10.1% for avoid search. For both faces and text in both tasks, these values are significantly lower than in the free-viewing task ($p < 10^{-5}$ for all, *t*-test) but are still higher than cell phones in the free-viewing task ($p < 0.001$, for all *t*-test). The fraction of cell phones fixated was not significantly different across the search tasks and the free-viewing task.

Furthermore, we compared the timing of the initial fixations (free search and avoid). For the face stimuli, we observed that in the free search task on-target saccade

Figure 6. (a) Saccade planning durations across the three stimulus categories. Comparison of duration of first fixation when subjects fixated on the high-level entities (upper panel—faces, middle—text, bottom—cell phones). Left two bars in each panel correspond to subjects' on-target fixations and the right two bars correspond to those that went off-target. Black bars comprise of trials where subjects were instructed that the search target could appear anywhere in the image. Blue bars represent avoid trials. Stars indicate significance for *t*-test comparison between rightmost bar and the three other bars in each panel (***$p < 0.001$, **$p < 0.02$). In the avoid task, fixations away from the face/text (represented by the rightmost bars in each panel) are significantly higher than each of the other three bars in that panel. (b) Fraction of first on-target fixations in the free search versus avoid search tasks across the three stimulus categories. Similar to free-viewing fixations, we show in each panel the fraction of trials in which a fixation landed on the ROI for the first time by the *n*th fixation.


Faces
***
350
300
250
200
First fixation on face
First fixation away from face
Text
**
First fixation on text
First fixation away from text
Cell phones
First fixation on phone
First fixation away from phone
Saccade planning time (ms)


Faces
Free search task
Avoid task
40
20
0
1
2
3
4
5
6
7
8
9
10
Text
Cell phones
Percent of trials
Fixation number

planning times took on average 248 ± 63 ms (Figure 6), while off-target fixations took 260 ± 75 ms. These differences are not significant. In the avoid task, on-target saccade planning times took 246 ± 73 ms on average, but off-target times took 297 ± 93 ms. Subjects in the avoid task take significantly longer to make their first saccade off-target compared to both the avoid task on-target ($p < 10^{-3}$, *t*-test) and the free search task off-target ($p < 10^{-4}$). Simply put, it takes longer to *not* look at a face when told to avoid it than to *not* look at a face when freely searching the scene. We also observe that when avoidance fails—i.e., when subjects look at a face under the avoid instruction—the time before the saccade initiates is not statistically different as when subjects look at a face under the free search instruction.

Looking at text, we observe a similar pattern in the timing of fixations. Again, the avoid task off-target timing (297 ± 106 ms) is significantly greater than the other three conditions—the avoid task on-target (261 ± 61 ms), the free search task off-target (275 ± 77 ms), and the free search task on-target (247 ± 52 ms) ($p < 0.02$ for all). There is no significant difference in the timing when the first fixation goes to a face versus to text. More so, like for faces, there is no significant difference between the timing of the on-target and off-target fixations in the free search task for text. Cell phones timing results show no significant differences between any of the four categories of fixations (Table 1).

In order to test whether the globally faster first fixations in the face block are due to faces attracting attention faster rather than a general adoption of a strategy during the "face" block, we further investigated how saccade planning times vary over the course of the experiment. We performed a linear regression relating the trial number to the saccade planning time throughout the course of each block. We found no regression coefficient that was significantly below 0, indicating that none of the entities (faces, text, or cell phones) shows a decrease in saccade planning time for any of the experiments (free view, free search, avoid search).

In line with the overall trend of cell phones having the least significant effect on fixations during search, performance is better for trials where the distracting entity is a cell phone. Subjects were able to correctly identify the presence/absence of the target in 85% ± 2% of the trials when the image contained a cell phone, while they could do so in only 82% ± 3% of trials containing a face, and 78% ± 3% of trials containing text. There is a significant difference in performance between faces and text as well as between text and cell phones ($p < 10^{-3}$ for both; *t*-test). In trials where subjects were told to avoid locations where target would not appear, performance decreased in all cases to 72% ± 13% for trials containing faces, 72% ±

	Faces	Text	Phones	
Saccade planning time (on-target)	203 ms	239 ms	313 ms	Free view
	220 ms	233 ms	299 ms	Free view (control)
	248 ms	247 ms	287 ms	Free search
	244 ms	250 ms	299 ms	Free search (control)
	246 ms	261 ms	273 ms	Avoid search
	241 ms	251 ms	292 ms	Avoid search (control)
Saccade planning time (off-target)	199 ms	241 ms	290 ms	Free view
	208 ms	236 ms	310 ms	Free view (control)
	260 ms	275 ms	284 ms	Free search
	260 ms	266 ms	310 ms	Free search (control)
	297 ms	297 ms	299 ms	Avoid search
	290 ms	282 ms	310 ms	Avoid search (control)
Fraction of trials where the ROI	89%	65%	8%	Free view
was visited within first 2 fixations	87%	53%	7%	Free view (control)
	24%	32%	12%	Free search
	33%	34%	6%	Free search (control)
	28%	37%	10%	Avoid search
	28%	29%	9%	Avoid search (control)
2-AFC accuracy	82%	78%	85%	Free search
	79%	78%	80%	Free search (control)
	72%	72%	82%	Avoid search
	69%	72%	79%	Avoid search (control)

Table 1. Summary of timing and accuracy results for the four experiments: Experiment 1—"free view," Experiment 2—"free view (control)," Experiment 3—"free search" and "avoid search," and Experiment 4—"free search (control)" and "avoid search (control)" and the three entities (faces, text, cell phones).

12% for trials containing text, and 82% ± 8% for trials containing cell phones. While the trend of significant difference ($p < 0.01$, t-test) between cell phones and the other two categories remains, the difference between faces and text is eliminated, supporting the claim that whether told to avoid looking at faces or text, subjects employ a similar search mechanism.

Experiment 4—control for the effect of adaptation

We further tested the absence of change in timing due to adaptation by measuring saccade planning time effects in a conjoint data set where effects of adaptation should not occur. This data set was created by stringing together the three types of entities in one long sequence as opposed to three separate blocks. The sequence consisted of both avoid and free search instructions. For the avoid task, we see a consistent trend of the data to that in the original experiment. For saccades going toward faces, the saccade planning time was 241 ± 60 ms, for text it was 251 ± 63, and for cell phones it was 292 ± 73 ms. Off-target timings also show the same increase in planning time for faces (290 ± 69 ms), text (282 ± 73), and even cell phones (310 ± 52 ms) ($p < 0.02$ for faces/text; t-test; cell phones' increase in timing is not significant).

We can also use these data to test whether subjects were changing their viewing strategies based on prior knowledge of the stimulus category. The free search set of intermixed trials is preceded by the same instruction regardless of stimulus entity, thus giving no information about the category of the stimulus prior to its presentation. We see that saccade planning time toward faces is 244 ± 69 ms, toward text is 250 ± 99 ms, and toward phones is 299 ± 68. The off-target timings are also consistent with the results of Experiment 3: 260 ± 60 ms for faces, 266 ± 62 for text, and 310 ± 65 for cell phones. This indeed shows that the timing effects are due to the biases these entities pose rather than a general effect of adaptation or category-dependent strategy. See Table 1 for a list of all performance values and the latency for each entity.

Model analysis

In order to improve the predictive performance of the saliency algorithm, we added specific channels dedicated to faces, text, and cell phones to the standard saliency map model. We calculated how well the new map predicts the locations of subjects' fixations.

The performance of the standard saliency map for viewing of images containing faces was on average 79.0% (normalized AUC). Adding the face channel increased predictability to 87.4%. Predictions of fixation location for subjects viewing text improved from 77.3% to 84.8%. Both increases in predictability are significant ($p < 10^{-6}$ for both, paired t-test). The text channel leads to improvements for every image containing text. For cell phone images, the mean AUC improved slightly but significantly (from 77.0% to 78.1%; $p < 10^{-3}$). Figure 7 compares the AUC for each of 231 individual images with the standard saliency map to the new saliency map.

Discussion

The results of the first experiment show that faces and text have a profound effect on the allocation of eye movements. The eyes are rapidly and strongly attracted to both, as compared to the slower and fewer fixations to cell phones (Figure 5). Faces and text rapidly attracted gaze—in over 65% of trials, faces and text were foveated within the first two fixations. In contrast, only 27.3% of cell phones were foveated even after seven fixations. Our results suggest similar mechanisms for attentional deployment to both faces and text, with a higher emphasis on faces. A second experiment controlling for the relative size and background of the images shows that faces are 1.49 times more likely to be attended to than text and 5.18 times more likely to be attended than cell phones.

In a third experiment, a separate group of subjects was asked to detect a small cross in the same natural scenes. The cross was present in half of the images. In three blocks, subjects were told before the trial began that the cross would not be present on the faces, text, or cell phones ("avoid"), while in the other three blocks, the cross could be found anywhere ("free search"). We reasoned that as subjects were under time pressure, they would not look at faces when they knew that the target was not located on faces, as this would be inefficient (same for text and cell phones). However, we see no such trend (Figure 6b and Table 1). There are no highly significant results discriminating between the avoid task and the free search task inferred from the fraction of fixations landing in the ROI. This was tested by comparing both the first one and the first two fixations of each trial. This observation may have not been seen due to the instructions, forcing the subjects to think of the targets to be avoided, and thus causing them to look at those targets more often (Wegner, 1994).

Our data shows a consistent relationship between saccade-planning time and the percentage of trials in which the region of interest was attended during the first fixation. In the free-viewing task, we see that across entities saccade planning time is fastest for faces, second fastest for text, and slowest for cell phones. This is consistent with the percentage of trials in which saccades landed on each of these entities' regions of interest, showing an inverse relationship to saccade-planning time. Further changes across categories also show this relationship. The fraction of trials in which faces and text are attended to initially is significantly lower in the free

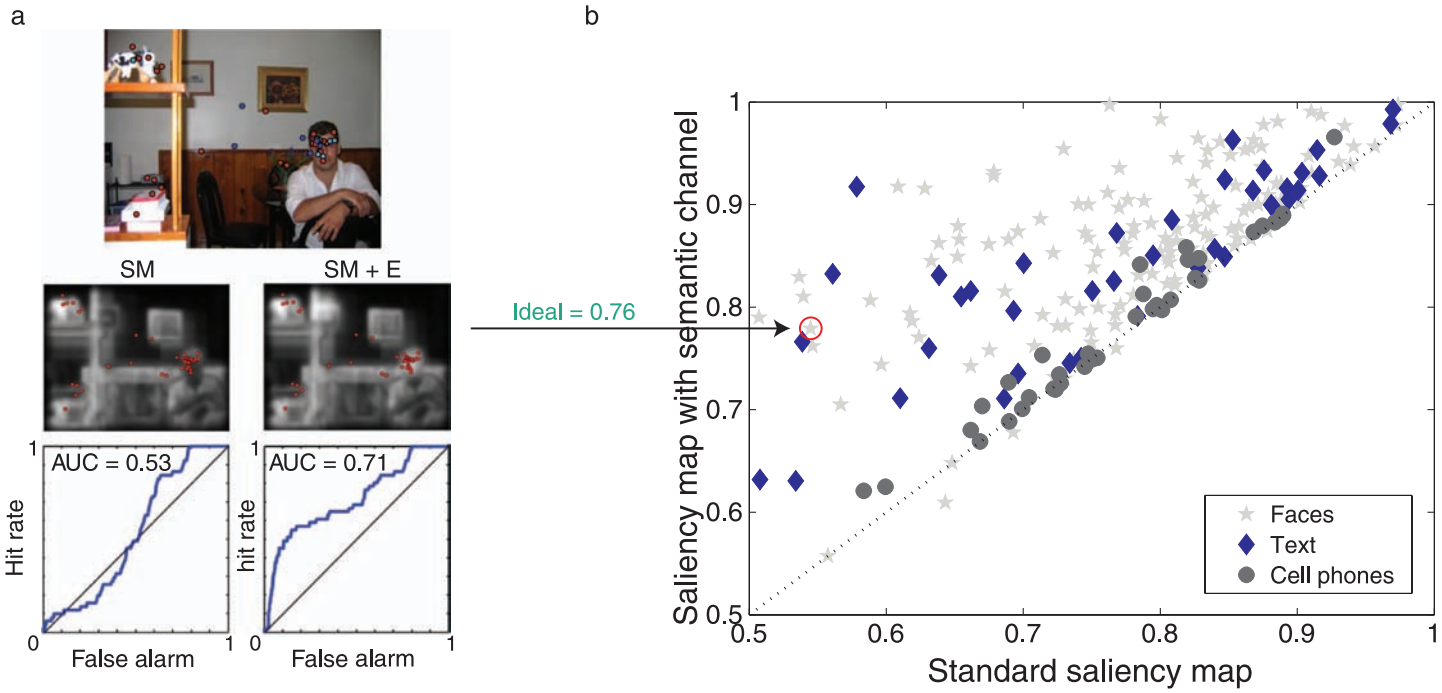


Figure 7. Performance comparison for all 231 images. (a) An example of the way by which each point in the scatter diagram was calculated. For the image pictured, the fixations of all subjects were superimposed and were compared using the ROC curve to both the standard saliency model (SM) and the saliency model with high-level entity channel (SM + E). The ROC curve is created by comparing the hit rate and false alarm rate for all possible thresholds. The AUC for each map is normalized by the ideal AUC. (b) Each symbol represents the model's performance predicting subjects' fixations on a particular image. Shapes and colors in the scatterplot indicate the different categories. Symbols above the diagonal indicate an improvement in the saliency map model with the inclusion of a high-level channel. Images with faces or text are improved by the addition of a high-level channel. Images with faces are improved the most while images with cell phone channel inclusion show only a marginal improvement.

search task than in the free-viewing task, this is reflected by the increase in saccade-planning time across the two tasks. For cell phones, there are no significant differences in saccade planning time across tasks, and we see no significant differences in the fraction of trials as well. We show that faces and text are significantly more salient than cell phones, by postulating that as the number of subjects who look at an ROI and the speed at which they look at it increases, the more salient it is regarded.

We report that it takes longer to initiate saccades for faces and text in the free search paradigm than in the free-viewing paradigm, but we see no such difference for cell phones. This indicates that the sensitivity to faces and text in the search paradigm has been depressed, resulting in slower saccade times ($p < 10^{-6}$ free view vs. free search faces; $p < 0.01$ free view vs. free search text), which is consistent with the decreased fraction of saccades landing in the regions of interest. This suggests that the salience of faces and text is task dependent. However, this sensitivity toward faces and text does not vanish completely, as saccade-planning time did not increase all the way to the level of cell phones ($p < 10^{-5}$ for both vs. cell phones; *t*-test). This suggests that attentional deployment to faces and text is regulated only partially by top–down mechanisms.

The avoid experiment gives further support to the claim that attention toward faces and text is in part reflexive (Bindemann et al., 2005). In the avoid task, subjects are better-off not looking at any of the entities; however, occasionally the subjects fail to avoid them. These are on-target saccades in the avoid task, and their saccade-planning time is not statistically different from on-target saccades in the free search paradigm. This indicates no decrease in sensitivity to faces and text, even when subjects are explicitly instructed to avoid the entity. In contrast, our data show that on average it took 37 ms longer for subjects to look *away* from a face in the avoid task than in the free search task. Similarly, it took 22 ms longer to look *away* from a text in the avoid task than in the free search task. This increase in saccade-planning time suggests that extra computational effort may be required to actively avoid looking at faces or text because there is a natural tendency to attend to these stimuli.

If attentional deployment to faces or text were a top–down process, saccades to these stimuli would likely require longer to initiate compared to saccades driven by a bottom–up process. However, we observe the planning of these saccades to be equivalent in duration in both search tasks. Additionally, if saccades to faces and text were top–

down, we would expect that avoiding these stimuli would not increase the time it takes to look away from these image elements. In fact, if the saliency of these elements could be quickly modulated by top–down mechanisms, we would expect to see a drop in the time it takes to avoid faces or text. Instead, we observe the opposite, with slower top–down driven computations preventing rapid deployment of bottom–up attention. More so, we see that when avoidance fails, saccade planning takes the normal amount of time. This suggests that top–down influences have not had enough time to influence attentional allocation, and thus bottom–up forces are mainly present. We show that while indeed faces can be avoided, there are aspects such as latency measures that still reflect the bias.

Further temporal analysis uncovered other facts (Table 1). Timing analysis of saccade initiation for the free-viewing task indicates that saccade initiation is fastest for faces, slower for text, and slowest for cell phones. These results go hand in hand with prior data reported by Fletcher-Watsonô, Findlay, Leekam, and Benson (2008), showing that latencies for saccades to faces during the first fixation in a free-viewing task were between 100 and 249 ms. Interestingly, off-target saccade times are indistinguishable from on-target saccade times in each of the image categories—seemingly implying that subjects look away from a face faster than they look away from text.

A possible explanation of the observed global depression in saccade latency could be that subjects may be adopting different general strategies across the block of images for each entity. To test this hypothesis, we ran a control experiment where the three entities were randomly intermixed in one long block. The saccade-planning times for the different entities were consistent with those observed in Experiment 3—indicating no adaptation across the course of the experiment. We pulled out the stimulus set that followed the "free search" instruction to be sure that the subjects had absolutely no knowledge of the stimulus category to come. Again, we found no significant differences between this set and our original experiments, thus showing this global depression phenomenon is not due to blocks or knowledge of the stimulus category. We also tested the possibility that this global depression is an image confound, as the images tested were disjoint sets across the stimulus categories. By embedding our entities into the same background images, we eliminated disparities in the image background that could be influencing the off-target saccades. However, even in this set of images, we still see this global depression of saccade latency based on stimulus category.

A more likely explanation can be seen by considering a possible statistical selection bias of off-target saccades. For a non-face object to be the target of a saccade, it must have a saliency level larger or close to that of the competing face target. Given, then, that a non-face target was attended to, the expected salience of this target will be higher if there is a highly salient face in the image than if there is not. Thus, during the face trials, only the high-saliency non-face objects are able to outcompete the faces for attention, and it is this subset used for the timing analysis. The fewer number of off-target saccades in the face trials select only these highly salient competitors. The larger number of off-target saccades in the text and phone trials also select out these highly salient competitors and more competitors that are less salient—thus bringing down the average. Under the stipulation that saccade planning time decreases with higher stimulus saliency, these stimuli will then have a faster saccade-planning time than for less salient stimuli (such as cell phones, and to a lesser extent, text, in the present study).

This selection bias also explains the global discrepancy in the artificially embedded image set. Consider two competing stimuli that are in the embedded image set, say a highly salient banana, and a moderately salient chair. When a face is embedded onto the image that contains the highly salient banana, the banana will win the competition and will draw a saccade with a fast saccade-planning time. When a face is embedded onto the image containing the moderately salient chair, the face will win the saccade. Thus, the saccade-planning time of off-target saccades will be the speed of the saccade going to the banana. When a text object is embedded onto the same images, both the banana and the chair will first attract a saccade. Therefore, the saccade-planning time of off-target saccades for text will be slower than for the face.

For further insight, imagine that the saliency of competing objects is drawn from a Gaussian distribution. If the competition with highly salient faces wins over the bottom 80% of these stimuli, then only the top 20% contribute to off-target saccade-planning times. If competition with less salient text items cuts off the bottom 60% of these stimuli, then the top 40% will contribute to the off-target saccade-planning times. This leads to overall slower off-target reaction times for images containing text (the average of the top 40%) compared to images containing faces (the average of the top 20%).

The early fixations to faces and text during free-viewing and search (even when told to avoid looking at faces and text) suggest that subjects are biased to look at these objects independent of top–down mechanisms that drive eye movements and covert attention (Honey, Kirchner, & VanRullen, 2008). The tendency to look at faces is so pronounced that it is even possible to decode which image is associated with which scanpath by using the exact location of faces in the image (Cerf, Harel, Huth, et al., 2008). This ability to decode which image corresponds to a given scanpath is a measure of how attractive various image features are, as they reflect lower across-subject variability in scanpaths.

Although the exact way in which attention to faces is implemented in the brain is unclear (Johnson, Dziurawiec, Ellis, & Morton, 1991), it is well known that a number of

cortical regions are specialized for faces, in particular the fusiform gyrus (Kanwisher, McDermott, & Chun, 1997; Tsao, Freiwald, Tootell, & Livingstone, 2006). In contrast to faces, we can be almost certain that attention to text is not an evolved process. Instead, it is likely that text sensitivity is developed through learning. It is possible, however, that the development of text itself was influenced by factors in the brain that control attention, and it is these factors that explain why text is highly salient (Changizi, Zhang, Ye, & Shimojo, 2006). Recent studies in fact argue in favor of a specialized area in the brain for words and text—the "visual word form area" (Cohen, Jobert, Le Bihan, & Dehaene, 2004), which could take part in the allocation of bottom–up attention demonstrated in our tasks. Interestingly, long-term experience may also play a role in development of face recognition abilities in the brain (Golarai et al., 2007). Regardless, our results show very similar patterns in the way by which attention is allocated toward faces and text, suggesting similar mechanisms for attention deployment in the brain.

Further studies of the interaction between top–down and bottom–up mechanisms leading to the attention allocation to faces show that the ability to rapidly saccade to faces in natural scenes depends, at least in part, on low-level information contained in the Fourier 2-D amplitude spectrum (Honey et al., 2008). This suggests that a bottom–up saliency model incorporating image features may be able to account for part of the attention allocation.

To formally evaluate the idea that faces and text are intrinsically salient, we compared predictions that the standard bottom–up saliency model makes for fixations in natural scenes. We used the predictions of the saliency algorithm enhanced by an additional map (Figure 2). The standard saliency model does not encode anything high-level or cognitive but only operates on center-surround maps defined at nine distinct scales for orientation, intensity, and color. The additional map encodes the location of all faces, text, and cell phones in the images. The saliency map analysis attempts to remove the contribution of low-level features from the salience of these entities, thus revealing only high-level features that contribute to gaze position. We see a large increase in performance when adding the face and text channels, suggesting that much of the salience of these features cannot be explained by their low-level features alone. Mainly, this can be shown by the lack of improvement for the cell phone channel, although it was also added to the saliency map in the same fashion.

Our experiments and our modeling demonstrate that faces and text are very attractive and are difficult to ignore, even if there is a real cost associated with looking at them. It remains to be seen how the single neuron substrate of faces and text reconciles the sometimes conflicting demands of bottom–up and top–down inputs. Our improved model of saliency-driven attentional deployment is of relevance to a host of military, commercial, and consumer applications. The success of incorporating additional biologically inspired detectors for high-level cues suggests similar attention allocation patterns to those used by the brain.

Acknowledgments

This research was supported by the National Institute for Mental Health and the Mathers Foundation. The authors wish to thank Kelsey Laird for valuable comments.

Author contributions: All authors contributed equally to this paper.
Commercial relationships: none.
Corresponding author: Moran Cerf.
Email: moran@klab.caltech.edu.
Address: 1200 California Boulevard, Pasadena, CA 91125, USA.

References

Bindemann, M., Burton, A., Hooge, I., Jenkins, R., & de Haan, E. (2005). Faces retain attention. *Psychonomic Bulletin & Review, 12,* 1048–1053. [PubMed] [Article]

Bindemann, M., Burton, A., Langton, S., Schweinberger, S., & Doherty, M. (2007). The control of attention to faces. *Journal of Vision, 7*(10):15, 1–8, http://journalofvision.org/7/10/15/, doi:10.1167/7.10.15. [PubMed] [Article]

Brainard, D. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. [PubMed]

Cashon, C., & Cohen, L. (2003). The construction, deconstruction, and reconstruction of infant face perception. In O. Pascalis & A. Slater (Eds.), *The development of face processing in infancy and early childhood: Current perspectives* (pp. 55–68). New York: NOVA Science Publishers.

Cerf, M., Cleary, D., Peters, R., Einhäuser, W., & Koch, C. (2007). Observers are consistent when rating image conspicuity. *Vision Research, 47,* 3052–3060. [PubMed]

Cerf, M., Frady, E., & Koch, C. (2008). Using semantic content as cues for better scanpath prediction. *Proceedings of the 2008 symposium on Eye tracking research & applications* (pp. 143–146). New York, NY, USA: ACM.

Cerf, M., Harel, J., Einhauser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems, 20,* 241–248.

Cerf, M., Harel, J., Huth, A., Einhauser, W., & Koch, C. (2008). Decoding what people see from where they look: Predicting visual stimuli from scanpaths. In L. Paletta & J. K. Tsotso (Eds.), *Lecture Notes in Artificial Intelligence, LNAI 5395* (pp. 15–26). Heidelberg: Springer-Verlag Berlin.

Changizi, M., Zhang, Q., Ye, H., & Shimojo, S. (2006). The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes. *American Naturalist, 167,* E117–E139. [PubMed]

Cohen, L., Jobert, A., Le Bihan, D., & Dehaene, S. (2004). Distinct unimodal and multimodal regions for word processing in the left temporal cortex. *Neuroimage, 23,* 1256–1270. [PubMed]

Cornelissen, F., Peters, E., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers, 34,* 613–617. [PubMed]

Dickinson, S., Christensen, H., Tsotsos, J., & Olofsson, G. (1994). Active object recognition integrating attention and viewpoint control. *Proceedings of the Third European Conference on Computer Vision, Stockholm, Sweden, May 2–6, 1994*.

Einhäuser, W., Kruse, W., Hoffmann, K., & König, P. (2006). Differences of monkey and human overt attention under natural conditions. *Vision Research, 46,* 1194–1209. [PubMed]

Einhäuser, W., Rutishauser, U., Frady, E., Nadler, S., König, P., & Koch, C. (2006). The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision, 6*(11):1, 1148–1158, http://journalofvision.org/6/11/1/, doi:10.1167/6.11.1. [PubMed] [Article]

Fletcher-Watsonô, S., Findlay, J., Leekam, S., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception, 37,* 571–583. [PubMed]

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8*(2):6, 1–17, http://journalofvision.org/8/2/6/, doi:10.1167/8.2.6. [PubMed] [Article]

Golarai, G., Ghahremani, D., Whitfield-Gabrieli, S., Reiss, A., Eberhardt, J., Gabrieli, J., et al. (2007). Differential development of high-level visual cortex correlates with category-specific recognition memory. *Nature Neuroscience, 10,* 512–522. [PubMed]

Honey, C., Kirchner, H., & VanRullen, R. (2008). Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. *Journal of Vision, 8*(12):9, 1–13, http://journalofvision.org/8/12/9/, doi:10.1167/8.12.9. [PubMed] [Article]

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506. [PubMed]

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews, Neuroscience, 2,* 194–204. [PubMed]

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,* 1254–1259.

James, W. (1890). *The principles of psychology*. New York: Holt.

Johnson, M., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition, 40,* 1–19. [PubMed]

Kanwisher, N., McDermott, J., & Chun, M. (1997). The Fusiform face area: A Module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17,* 4302–4311. [PubMed] [Article]

Mack, A., Pappas, Z., Silverman, M., & Gay, R. (2002). What we see: Inattention and the capture of attention by meaning. *Consciousness and Cognition, 11,* 488–506. [PubMed]

Oliva, A., Torralba, A., Castelhano, M., & Henderson, J. (2003). Top–down control of visual attention in object detection. *Proceedings of the 2003 International Conference on Image Processing* (p. 1).

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107–123. [PubMed]

Peters, R., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom–up gaze allocation in natural images. *Vision Research, 45,* 2397–2416. [PubMed]

Rizzolatti, G., Riggio, L., Dascola, I., & Umilta, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia, 25,* 31–40. [PubMed]

Ro, T., Russell, C., & Lavie, N. (2001). Changing faces: A detection advantage in the flicker paradigm. *Psychological Science, 12,* 94–99. [PubMed]

Simion, C., & Shimojo, S. (2006). Early interactions between orienting, visual sampling and decision making in facial preference. *Vision Research, 46,* 3331–3335. [PubMed]

Tatler, B., Baddeley, R., & Gilchrist, I. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45,* 643–659. [PubMed]

Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition, 13,* 657–665.

Tsao, D., Freiwald, W., Tootell, R., & Livingstone, M. (2006). A cortical region consisting entirely of face-selective cells. *Science, 311,* 670–674. [PubMed] [Article]

Vuilleumier, P. (2000). Faces call for attention: Evidence from patients with visual extinction. *Neuropsychologia, 38,* 693–700. [PubMed] [Article]

Wegner, D. (1994). *White bears and other unwanted thoughts*. New York, NY, USA: Guilford Press.