PERGAMON

Vision Research 43 (2003) 333–346

Vision Research

www.elsevier.com/locate/visres

Oculomotor strategies for the direction of gaze tested with a real-world activity

Kathleen A. Turano *, Duane R. Geruschat, Frank H. Baker

Wilmer Eye Institute, The Johns Hopkins University School of Medicine, Baltimore, MD 21205-2020, USA

Received 12 November 2001; received in revised form 17 June 2002

Abstract

Laboratory-based models of oculomotor strategy that differ in the amount and type of top-down information were evaluated against a baseline case of random scanning for predicting the gaze patterns of subjects performing a real-world activity—walking to a target. Images of four subjects' eyes and field of view were simultaneously recorded as they performed the mobility task. Offline analyses generated movies of the eye on scene and a categorization scheme was used to classify the locations of the fixations. Frames from each subject's eye-on-scene movie served as input to the models, and the location of each model's predicted fixations was classified using the same categorization scheme.

The results showed that models with no top-down information (visual salience model) or with only coarse feature information performed no better than a random scanner; the models' ordered fixation locations (gaze pattern) matched less than a quarter of the subjects' gaze patterns. A model that used only geographic information outperformed the random scanner and matched approximately a third of the gaze patterns. The best performance was obtained from an oculomotor strategy that used both coarse feature and geographic information, matching nearly half the gaze patterns (48%). Thus, a model that uses top-down information about a target's coarse features and general vicinity does a fairly good job predicting fixation behavior, but it does not fully specify the gaze pattern of a subject walking to a target. Additional information is required, perhaps in the form of finer feature information or knowledge of a task's procedure.

Keywords: Oculomotor search strategies; Mobility; Gaze; Visual saliency; Guided search

1. Introduction

While engaged in most activities, "the eye moves in a series of quick jerks and pauses" (Buswell, 1935). The "quick jerks", or saccadic eye movements, occur about three to four times a second. These eye movements redirect the retinal area capable of receiving the most finely detailed information (typically the fovea) from place to place to obtain information with high spatial resolution. The "pauses" or fixations are thought to minimize image blur and allow the visual system time for processing the image. The factors that determine what parts of the image to fixate or attend [1] has been a matter of controversy; some claim that the choice of fixation location is best described as random (Kundel, Nodine, Thickman, & Toto, 1987), others suggest that stimulus factors are critical determinants (Geisler & Chou, 1995; Itti & Koch, 2000; Itti, Koch, & Niebur, 1998; Niebur & Koch, 1996; Parkhurst, Law, & Niebur, 2002; Toet, Kooi, Bijl, & Valeton, 1998), still others claim that cognitive factors or expectations play a key role (Land & Horwood, 1995; Land & McLeod, 2000;

* Corresponding author. Address: Lions Vision Center, 550 N. Broadway, 6th floor, Baltimore, MD 21205, USA. Tel.: +1-410-502-6434; fax: +1-410-955-1829.

E-mail address: kathy@lions.med.jhu.edu (K.A. Turano).

[1] While it is known that one can shift attention without shifting gaze, visual attention typically precedes a saccade to the same location (Inhoff, Pollatsek, Posner, & Rayner, 1989; Kowler, Anderson, Dosher, & Blaser, 1995; Shepherd, Findlay, & Hockey, 1986). Electrophysiological data indicate that the two share some of the same neurophysiology (Posner & Petersen, 1990) supporting the view that the two are tightly linked (Corbetta, 1998; Kustov & Robinson, 1996).

0042-6989/02/$ - see front matter
PII: S0042-6989(02)00498-4

Land, Mennie, & Rusted, 1999; Land & Lee, 1994; Wolfe, Cave, & Franzel, 1989).

Until 1935, when Buswell published his systematic study of where people direct their eyes (or gaze) while looking at pictures, direction of gaze had been studied almost exclusively with the task of reading (Buswell, 1935). The change from reading to picture viewing introduced a new task and another significant dimension (verticality) to the study of gaze direction. More recently gaze patterns have been obtained as people engage in more physical activities such as driving, (Kito, Haraguchi, Funatsu, Sato, & Kondo, 1989; Land & Horwood, 1995; Land, 1992; Land, 1998; Land & Lee, 1994; Wann & Swapp, 2000) tea making, (Land et al., 1999) and cricket playing (Land & McLeod, 2000). These tasks have added yet more complexity to the stimulus. Still, the development of corresponding models of oculomotor strategies has failed to keep pace with the newly emerging descriptive studies.

The models of oculomotor strategies that exist have been derived from laboratory-based studies of visual search (Kundel et al., 1987; Treisman & Gelade, 1980; Wolfe et al., 1989) and can differ in the amount of emphasis placed on bottom-up (stimulus dependent) or top-down (task dependent) factors. One bottom-up oculomotor strategy, the visual salience model, is based on the premise that a person directs his or her gaze at the most visually salient location in the retinal image (Itti & Koch, 2000; Itti et al., 1998; Koch & Ullman, 1985; Theeuwes & Burger, 1998). Common to this strategy is the concept of a topographically organized "saliency map", which assigns a visual salience value to each point in the image. The saliency map is similar to the "master map" of Treisman and Gelade's (1980) which is derived from the integration of various feature maps (intensity, color, orientation). The visual salience strategy receives support from studies that have compared human performance against a computational model of visual salience (Itti & Koch, 2000; Itti et al., 1998). Itti et al.'s visual salience model allocates attention according to the rank order of the visual salience in an image. Their model is able to predict performance on pop-out attention tasks. Using the same model to compute the visual salience of fixation locations of subjects free viewing pictures of scenes, Parkhurst et al. found a higher visual salience value for the fixation locations than a calculated visual salience value from randomly sampled image locations (Parkhurst et al., 2002).

Unlike the bottom-up, visual salience model, the guided search oculomotor strategy is based on the idea that information about the nature of the target differentially weights specific features and can bias the direction of gaze. With guided search models, (Hoffman, 1978, 1979; Neisser, 1967; Wolfe et al., 1989) display items are evaluated according to their similarity to the expected target in an initial stage of parallel processing. Items that are similar are considered candidate targets and are passed onto a serial comparison stage for closer inspection (using selective attention and gaze). In this model, attention (and gaze) is directed to the item with the highest similarity value with subsequent shifts to items of decreasing similarity.

While the laboratory allows control over many extraneous factors as well as the opportunity to manipulate variables of choice, the ultimate goal in behavioral research is to apply laboratory-based knowledge to performance in the real world. Oculomotor strategies have not been formally tested in the few real-world eye-movement studies that have been conducted. But the tight coupling between gaze and task-relevant information in some studies suggests that top-down information may play a key role in guiding fixation (Hayhoe, Bensinger, & Ballard, 1998; Land et al., 1999; Land & Lee, 1994). For example, in a driving study, subjects directed their gaze 80% of the time to a specific geographic location when entering a bend in the road, i.e., the tangent point of the curve (Land & Lee, 1994). And in a tea-making task, subjects fixated specific items relevant to the task, e.g., teakettle and cup (Land et al., 1999). If top-down information does play a role in directing gaze in real-world tasks, then the question arises as to what kind of top-down information is used.

In this study we developed a method for quantifying gaze patterns in a real-world task to allow testing of various oculomotor models. We evaluated oculomotor strategies, which differed in the amount and type of top-down information used to guide fixation, against a baseline case of random scanning. Four oculomotor strategies were tested: no top-down information (visual salience model), information about the target's features (feature model), information about the general vicinity of the target (geographic model), information about both the target's features and general vicinity (feature–geographic model). Walking to a target was chosen as the real-world task. It is commonly performed in daily living, has a well-defined target, and falls somewhere between free viewing and driving in terms of attention constraints.

2. Methods

2.1. Subjects

We tested four visually normal persons. Their binocular visual acuity, corrected if necessary, measured with a Lighthouse ETDRS acuity chart (Ferris, Kassoff, Bresnick, & Bailey, 1982), was better than 20/25. Their binocular log peak contrast sensitivity, measured with the Pelli–Robson chart (Pelli, Robson, & Wilkens, 1988), was better than 1.65. Table 1 lists the ages, visual function measures, and travel times of the subjects.

Table 1
Subject characteristics

Subject	Age (years)	Visual acuity	Log MAR	Log CS	Travel time (s)
NPD	49.4	20/14	−0.16	1.7	21.3
NJF	57.8	20/18	−0.06	1.9	27.2
NEL	66.2	20/22	0.04	1.7	23.8
NLT	36.2	20/15	−0.12	1.9	19.3

2.2. Mobility task

The mobility route consisted of the corridors of a floor in an office building that had never been seen by any of the subjects. An experimenter followed the subject throughout the route and recited standardized directions at specified points along the way. Each subject was instructed to walk safely, at his or her normal pace, following the directions given. The directions for the section of the route analyzed for this study were "As you walk down this hall, find the fifth door on the left and turn to go through". The distance for this section of the route was 24.8 m. Fig. 1 shows a picture of the hallway, indicating the target with an arrow.

Fig. 1. Picture of the hallway segment of the mobility route analyzed for this study. Instructions given to the subject were "As you walk down this hall, find the fifth door on the left and turn to go through". The numbers on the picture indicate the doorways on the left, and the arrow indicates the location of the target, the fifth door on the left.

2.3. Apparatus to record eye and scene during mobility

Images of eye and scene were recorded with an ISCAN (ETL-410) non-invasive, headband-mounted, eye and scene video-based imaging system, modified to be battery operated (Fig. 2). The system is lightweight (total weight 15 oz) consisting of a headband with two cameras, lenses, and a beam splitter. One camera images the subject's right eye and another camera (with a wide lens—$88° \times 60°$ field of view) images the scene. The camera outputs were recorded on digital video camcorders (Canon ZR10) carried in a backpack and analyzed offline. The camcorders were synchronized by recording a tone on the audio channels of the two camcorders, simultaneously. Prior to mounting the eyetracker headband, a silicon swim cap was fit on the subject's head to ensure positional stability.

2.4. Procedure to record eye and scene during mobility

The eyetracker was calibrated before the mobility data were collected. The subject was seated, placed on a bitebar, and instructed to fixate sequentially each of five points of a calibration pattern positioned at a distance of 1.3 m. The recorded values were used in the offline analysis (see below) to re-locate ISCAN eye-position values to direction of gaze. Upon completion of the mobility data collection, the eye recording was fed into

Fig. 2. Recording apparatus mounted on a person. Images of eye and scene were recorded with an ISCAN headband-mounted, eye and line-of-sight scene video-based imaging system. The system was modified to be battery operated. The cameras outputs were recorded on synchronized, digital video camcorders carried in a backpack and analyzed offline.

the ISCAN processing board that was externally triggered by the synchronizing tone. The ISCAN software uses the pupil and corneal reflection to identify the angular position of the eye. The scene recording used a video capture board (Broadway by Data Translation) whose software was modified to trigger on the synchronization tone on the videotape. In-house software was developed to transform the eye position data in ISCAN units to screen coordinates. Software was also developed to adjust eye position in accordance with the barrel distortion of the image introduced by the scene camera of the ISCAN system, as is clearly seen in Fig. 6. We used a lookup table based on the actual measured degree of distortion across the image. Movies of the eye-on-scene were made for each subject, which a graphic character was superimposed on each frame of the movie at the calculated eye position.

Fixations were identified using a velocity threshold of eye position relative to a scene landmark. To do this, for each frame of the scene movie, a distant stationary landmark was digitized and its coordinates stored. The change in the distance between the eye and the landmark was computed across consecutive frames. A fixation was defined as the eye-on-scene position remaining within 1.6° on two frames, equivalent to a velocity threshold of 24°/s. (It should be pointed out that the method of analyzing eye position from images stored on videotape, where the sampling rate is 30 frames/s, produced a temporal averaging of the eye position signal.) A criterion value of 1.6° was chosen as a compromise between the values of saccade detection used by Epelboim et al. (1997) (position change $> 2°$) and Zelinsky and Sheinberg (1997) (saccade onset velocity $> 20°$/s). Fixations detected using our criterion were initially cross-checked by visual inspection of the eye-on-scene movies. The good agreement between fixations detected by the two procedures led us to adopt the 1.6° criterion value. [2] The frames of the scene movie that correspond to each identified fixation will be referred to as "fixation frames". By definition, each fixation is associated with a minimum of two fixation frames.

[2] Image expansion during forward motion introduces a change between the eye and landmark. The magnitude of change is a function of the speed of the observer and the position of the environmental element. Elements that are eccentric to the observer's direction of motion and are close up introduce the most change with forward motion. In our study where average forward motion was 1.1 m/s, image positions of eye $> 15°$ from the center of the display will produce a change greater than the 1.6° threshold criterion if the underlying environmental element is closer than 2 m from the observer. A quick view of the movies suggests that very few eye positions were on elements this close. However, to ensure that no actual fixations were bypassed as a consequence of forward-motion changes, we visually rechecked the movie frames that corresponded to image positions of eye $> 15°$ from display center. For each subject the number of frames was fewer than 10%. The visual inspection resulted in no additional fixations being identified.

2.5. Models implementation

The stimulus for the subjects in our study was a temporally varying view of the environment. This view was sampled in time (30 frames/s) and space ($88° \times 60°$ image at 4 pixels/°) and stored on videotape for later analyses. The models that we tested were developed to process static images. In an attempt to provide the models with information most similar to what was available to the subjects, we used the fixation frames of each subject as input to the models. (Only one frame per fixation was used, the initial frame.) This approach eliminated any potential difference in temporal sampling between model and subject and resulted in the same number of predicted fixation locations across models for each subject.

To generate the predictions for the baseline case of random scanning, we implemented a *random scanner* loosely based on that of Kundel et al. (1987). The scanner randomly selected the x and y coordinates for each fixation location. The scanner had no memory for past fixations, and, therefore, multiple fixations could be made to the same location. Two versions of the random scanner were tested. In the "totally random" version, the scanner randomly selected x and y coordinates from anywhere on the image. In the "realistic" version, the scanner randomly selected the direction of the next fixation from one of 360°, and the distance to the next fixation was randomly selected from a probability density function of real eye movements (shown in Fig. 3). The function was an exponential with a mean of 4.4°.

For the *visual salience model*, we used a computer implementation (Itti & Koch, 2000) of a model developed by Itti et al. (1998) (This model is described in detail in Itti et al., 1998). (Details specific to our implementation are explicitly stated in this section.) This model, derived from the hypotheses and concepts proposed by Koch and Ullman (1985), is related to Treisman and Gelade's (1980) "feature integration theory" of attention. The input to the model is a digital color image that is filtered and progressively sub-sampled in a Gaussian pyramid scheme (Adelson, Anderson, Bergen, Burt, & Ogden, 1984; Burt & Adelson, 1983) to produce nine spatial scales of the image. Various feature maps are computed by a set of center-surround operations performed across spatial scales (i.e., the difference between a fine and a coarse scale). Based on the properties of the neural mechanisms of the primate visual cortex, the features are intensity, color (red–green, blue–yellow) and orientation (0°, 45°, 90°, 135°). Calculations were made at six center-surround combinations yielding 42 feature maps: 6 for intensity, 12 for color and 24 for orientation. The maps encode the local feature contrasts at various combinations of center and surround scales. A parameter file can be invoked to bias or differentially weight the coefficients for the feature maps. In the sa-

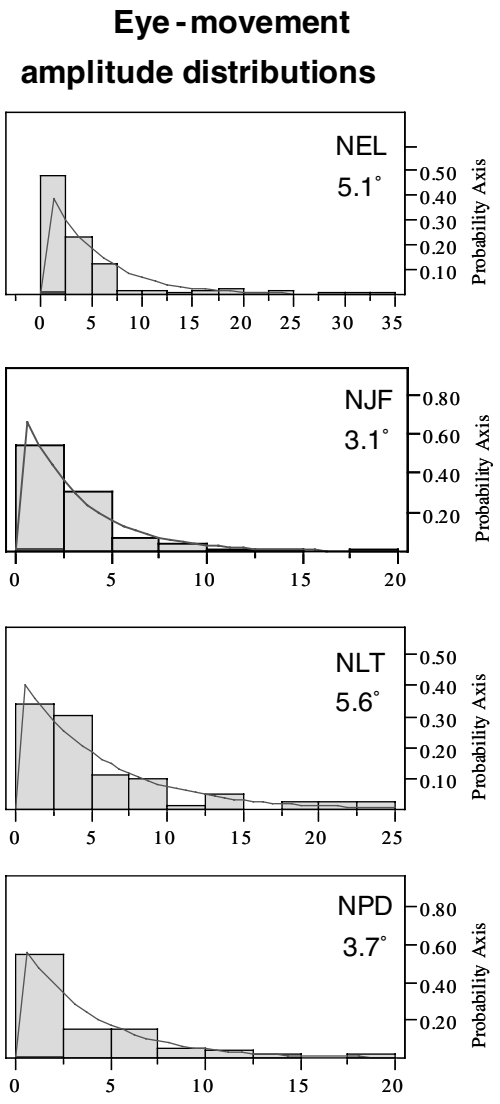

Fig. 3. Probability density functions of the eye-movement amplitudes (in degrees of visual angle) for the subjects. The curves represent the best-fit exponential functions. Means are indicated in the upper right corner of each graph.

liency model, all coefficients are 1. Each feature map is normalized, summed across scales, and the three resulting maps (intensity, color, and orientation) are summed to create a single two-dimensional saliency map. The most salient location, found by a winner-take-all algorithm, determines the location of the next fixation. To prevent permanent fixating on the most salient location the model includes an "inhibition of return" (IOR) component. Once a location has been fixated, that location is subsequently inhibited, and the next fixation shifts to the next most salient location. In our study we manually implemented this property since the model did not keep track of its predicted locations across input images. We visually inspected each predicted location, checking it against a log of previously predicted locations. If the location had been previously chosen, it was eliminated and the model chose the next most salient location for its prediction. The model also incorporates a "proximity preference", that is, when two locations are similar in salience, the closer location is chosen for the next fixation. Fig. 4 shows a sample image (Fig. 4a) and the corresponding saliency map (Fig. 4b). The location of the highest visual salience, and therefore the predicted fixation location, is represented as a small square centered in the circle in Fig. 4c.

The *feature model* is based on the idea that information about the nature of the target differentially weights specific features and can bias the direction of gaze. In the present study, the task was "to find the fifth door on the left and turn to go through", making the fifth door on the left the target. To implement the feature model, we used the visual salience model described above and modified the parameter file to maximize the weights of the feature map that codes for "vertical" and "large". To do this, we set to 1 the coefficient for the 0° orientation (vertical) feature at the center-surround combination with the lowest sub-sampled center and intermediate sub-sampled surround. The coefficients for the other orientations and center-surround combinations were set to 0. We eliminated the IOR component of the visual salience model. (This was done to be consistent with current guided search models.) No other changes were made to the visual salience computer program.

The *geographic model* is based on the idea that information about the general vicinity of the target biases the spatial location of fixation. To implement the geographic model for the current task, we applied the visual salience model described above, restricted fixations to the left side of the image, and eliminated the IOR component.

The *feature–geographic model* is a combination of the feature model and the geographic model. With this model, information about both the target's features and the general vicinity of the target are used to guide fixation. To implement this model we used the feature model described above but restricted fixations to the left side of the image.

2.6. Analysis

A categorical analysis, similar to that used by Stark and colleagues (Choi, Mosley, & Stark, 1995; Stark & Choi, 1996), was used to analyze the data. We chose this type of analysis instead of a pixel-based (x, y, coordinates) scheme because of our interest in the scene elements that subjects fixate (e.g. doors, posters, floor) rather than the pixel locations on the image. The distance between the actual eye position and a model's prediction may be only a few pixels, yet the eye and model prediction may be directed at different objects. For example, if the eye is directed at the bottom corner

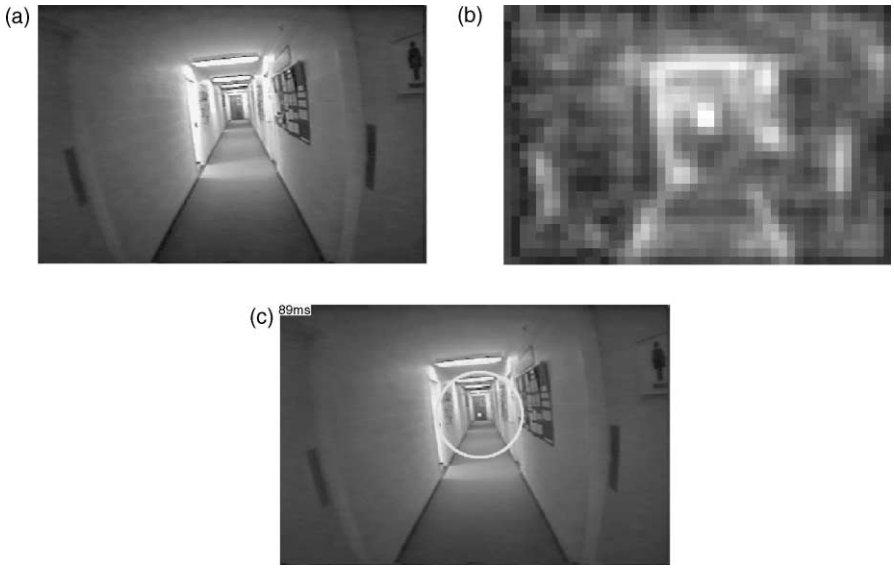

Fig. 4. Implementation of the visual salience model. (a) A grayscale version of a sample input image to the model. (b) Feature maps (intensity, color, and orientation) are computed by a set of center-surround operations performed across spatial scales. Each feature map is normalized, summed across scales, and the resulting maps are summed to create a single two-dimensional saliency map. (c) The most salient location determines the location of the next fixation (a small square centered in the circle).

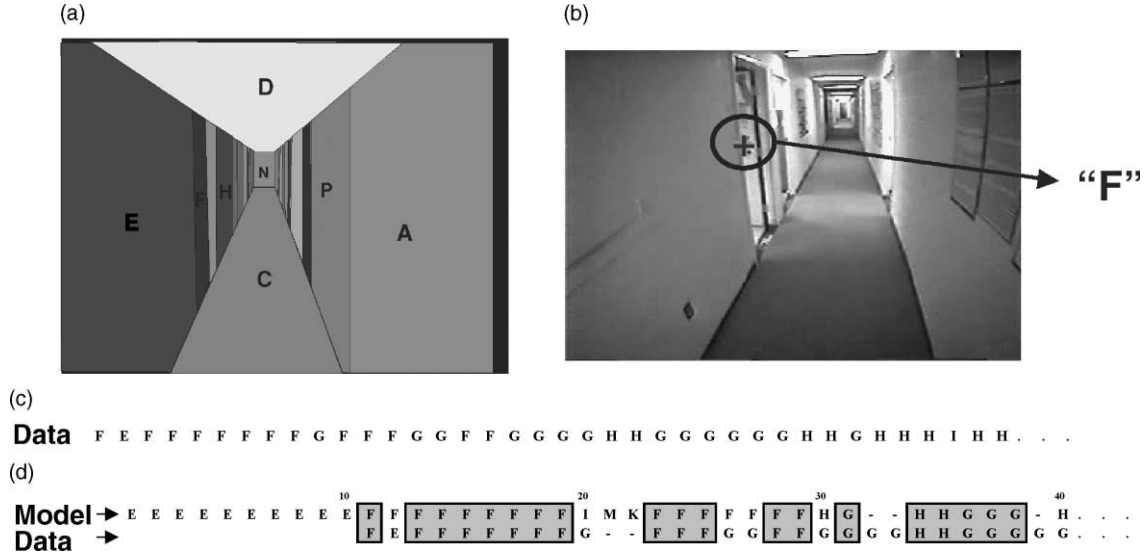

Fig. 5. Illustration of steps in analysis. (a) Caricature to illustrate various scene categories. (b) Classification of fixations. (c) Sample of fixation data string. (d) Portion of a sample sequence alignment (dashes indicate inserted spaces to achieve optimal alignment).

of a door and the predicted location is on the floor nearby, a pixel-based scheme would rate the similarity between the two high whereas a categorical type of analysis (using an object classification scheme) would rate the similarity low. In our study the categories were defined according to meaningful partitions. The images of the route were divided into 20 categories (e.g. first door on the left, wall between the first and second doors on the right, floor, ceiling). Each category was assigned a letter. Fig. 5a shows a cartoon of a scene illustrating some of the categories.

For each fixation, the position of the eye relative to the scene was classified into one of the 20 categories, using the video frame at the beginning of each fixation (Fig. 5b). A completed classification for each person was a string of letters representing the sequence of fixation locations (Fig. 5c).

For each model, the predicted location for each fixation was classified into one of the 20 pre-defined categories in the same manner as described above for the actual eye data. A sequence alignment analysis, designed for protein analysis (CLUSTALW from the MACVec-

tor software by Oxford Scientific), was used to determine the optimal alignment of the model predictions and the data. The parameters of the algorithm were set for maximizing alignment while minimizing the number of gaps. The algorithm used a residue weight matrix (BLOSUM series) and incorporated penalties for opening gaps (penalty set at 10) and extending gaps (set at 0.05). The gap-opening penalty gives the cost of opening a new gap of any length and the gap-extension penalty gives the cost of every item in a gap. Since it is usually difficult to align sequences that are most different from all other sequences, the program delayed the alignment of sequences that were less than 40% identical to any other sequence until all other sequences were aligned. The number of matched pairs was determined from the alignment results and the percent similarity between the two gaze patterns was calculated. Fig. 5d illustrates a portion of a sample alignment. Gray boxes indicate the matched pairs. To minimize selection bias from the computation of the random scanning, we ran the two random scanners on the fixation frames of each subject 10 times and calculated the optimal alignment for each of the samples. The mean of the percent similarity scores served as the final similarity score.

3. Results

To obtain an estimate of the magnitude of error in our method of positioning eye on scene we measured the distance between known image coordinates and calculated eye-on-scene positions as subjects looked at points in the world. The results showed that the average error of eye-on-scene locations was less than 0.5° for screen positions as far as 37° from scene center. To illustrate the relative size of possible position error relative to the scene Fig. 6 shows a video frame with the fixation marked by a black square at the arrow flanked by white squares that extend out by 0.5° on each side.

Fig. 6. Precision of fixation measurements. A video frame with the fixation marked by a black square at the arrow flanked by white squares that extend out by 0.5° on each side.

Fig. 7 shows pictures of the scene with superimposed alphanumeric characters to indicate the fixation locations of the subjects. The left panel shows the data collected in the first 7.5 s, and the right panel, data collected between 7.5 and 15.0 s. Data collected beyond 15 s are counted in the analysis but not displayed here. The complete eye-on-scene recordings can be seen by viewing the movies at the Web site, http://162.129.125.249/gaze.html. In the movies, a red cross indicates the eye position and a blue cross indicates the occurrence of a blink.

A comparison across subjects illustrates several common characteristics within the group. One, most fixations (71%–96%, mean = 83%) were on the side of the scene that contained the target, i.e., the left side. Two, most fixations (59%–82%, mean = 69%) were on the doors on the left side. Three, approximately two-thirds of the fixations (53%–72%, mean 62%) were on previously fixated left-side doors.

The degree to which the models' predicted gaze patterns matched the subjects' gaze patterns is shown in Table 2. The similarity scores were determined from the sequence alignment analyses. The two leftmost data columns show the similarity scores for the random scanners. The mean scores for the "totally random" and the "realistic" versions were 22.2% and 23.3%, respectively. (Since the similarity scores for the two random scanners were comparable, further references to analyses with the random scanner pertain to the realistic version.) The mean similarity scores for the visual salience model and the feature model were 21.3% and 20.8%, respectively. Both models generated gaze patterns that were no more similar to the subjects' gaze patterns than those generated by a random scanning procedure. On the contrary, the models that incorporated information about the general vicinity of the target, i.e., the geographic model and the feature–geographic model, generated gaze patterns that were more similar to the subjects' gaze patterns than that generated by the random scanner. The mean similarity score for the geographic model was 33.8%, significantly higher than the similarity score of the random scanner, $t(3) = 5.42$, $p < 0.01$. The mean similarity score of the feature–geographic model was 47.5%, also significantly higher than the score of the random scanner, $t(3) = 6.8$, $p < 0.01$.

In the above section we discussed the similarity between the models and data with respect to the *sequences* of fixations, or gaze patterns. In that analysis, the order in which the fixations were executed mattered. To determine whether the models predict where the subjects looked, regardless of order, we can compare the distributions of actual fixations and the models' predicted fixations. Distributions are shown for the subjects' fixations (Fig. 8a), and the models' predictions. The labels on the horizontal axis are the scene categories with the labels on the lower axis abbreviated category names and

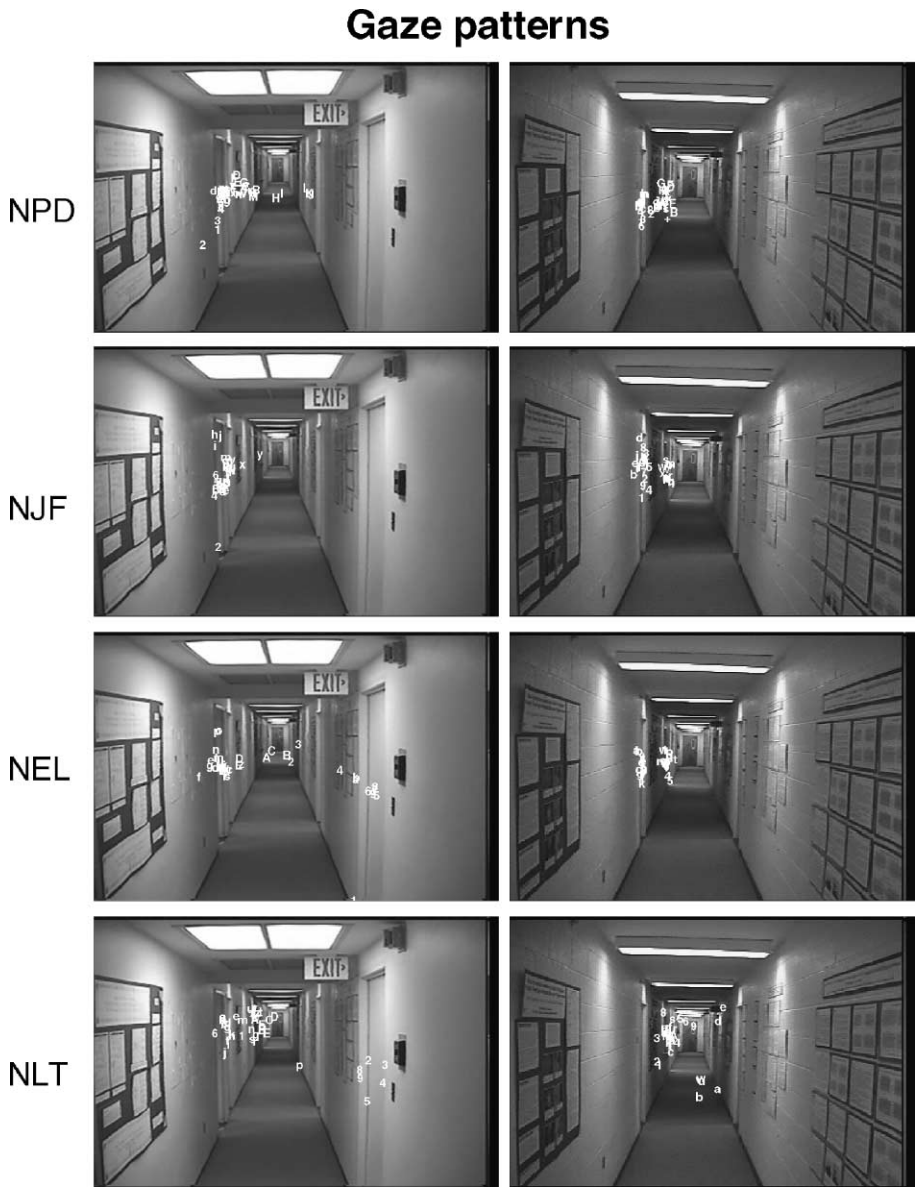

Fig. 7. Pictures of the scene with superimposed alphanumeric characters illustrating the eye-on-scene locations for the fixations of the subjects. The left panel shows the data collected in the first 7.5 s, and the right panel, data collected between 7.5 and 15.0 s. Fixation order is coded by the sequence: 1–9, a–z, A–Z.

Table 2
Similarity scores

	Random scanning		Visual salience (%)	Feature model (%)	Geographic model (%)	Feature–geographic (%)
	Totally random (%)	Realistic (%)				
NEL	24	23	26	22	28	48
NJF	18	22	16	14	34	42
NLT	18	20	16	17	34	38
NPD	29	28	27	30	39	62
Mean	22.2	23.3	21.3	20.8	33.8	47.5

Percent similarity between the models' predictions and subjects' data.

the labels on the upper axis in Fig. 8a arbitrary codes to serve as keys for labels in Figs. 5a and 9. Multiple occurrences of a category are coded by an "L" or "R" suffix, to indicate side of scene, followed by a number to

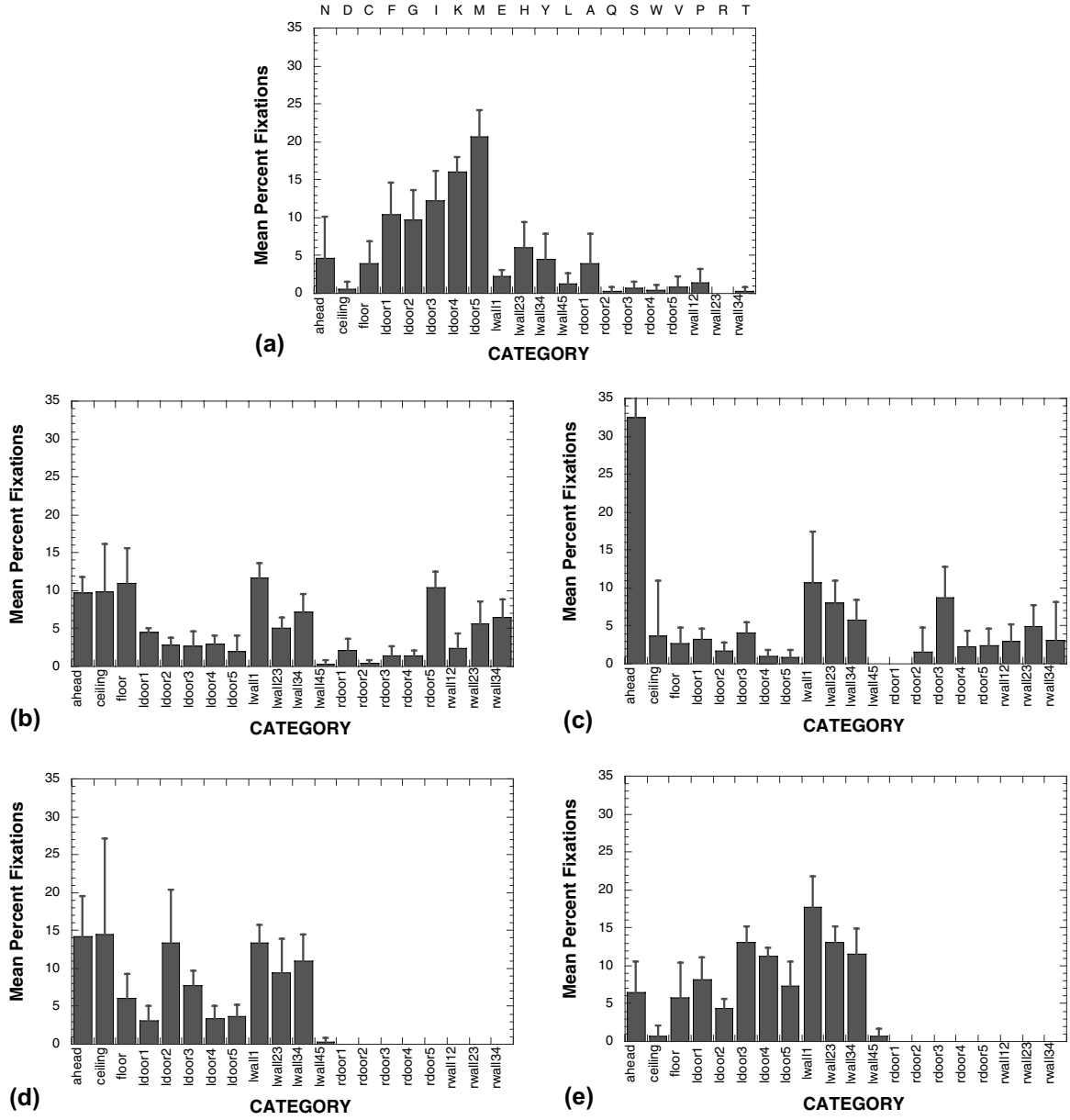

Fig. 8. Fixation percentages for subject data and four models. Frequency distributions of (a) actual fixations, and predicted fixations from (b) the visual salience model, (c) the feature model, (d) the geographic model, and (e) the feature–geographic model for the various categories. Labels on the lower *x*-axes are abbreviated category names. Multiple occurrences of a category are coded by an "L" or "R" suffix, to indicate side of scene, followed by a number to indicate order of occurrence relative to the beginning of the route. Labels on the upper *x*-axis are arbitrary category codes that serve as keys.

indicate order of occurrence relative to the beginning of the route. For example, "doorL1" indicates the first left-side door and "wallL23" indicates the wall on the left side between the second and third doors. (The wall categories include any existing posters on the walls.)

As shown in Fig. 8a, the left-side doors are fixated most often, and with the exception of the first door on the right (doorR1), the right side and ceiling are fixated the least. Fig. 8b shows the distribution of the visual salience model predictions. The distribution is more evenly apportioned across the categories compared to the distribution of the actual fixations (Fig. 8a). The variability in the model predictions (depicted by the size of the error bars) results from the different input images of the four subjects. Changes in head position as well as differences in the timing of fixations along the route cause different images in the camera's field of view across subjects.

The visual-salience model predicted that the categories: ahead, ceiling, floor, nearest left wall (wallL1), and the fifth door on the right (doorR5) have the most fixations. Over half of the fixations were predicted to fall

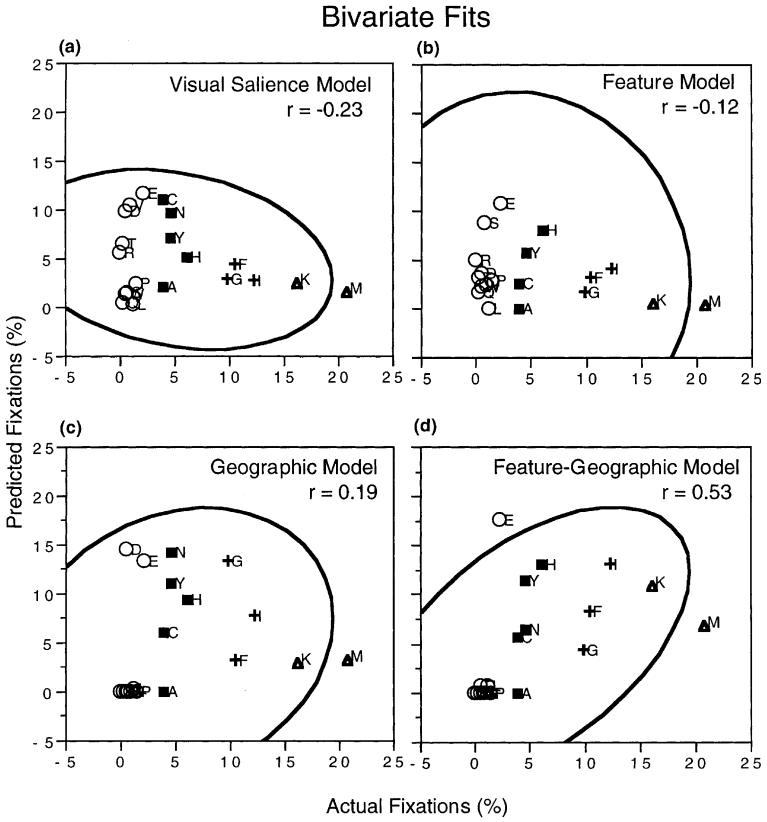

Fig. 9. Bivariate fits of predicted and actual fixations. Predicted frequencies from (a) the visual salience model, (b) the feature model, (c) the geographic model, and the (d) feature–geographic model plotted against the actual fixation percentages for the various categories. Bivariate normal density ellipses show where 95% of the data are expected to lie. The correlations between the predicted percentages of fixations of the models and the actual frequencies are shown in the upper right corners of each graph. Each datum is coded by a cluster symbol and a category code (the key for the category codes is in Fig. 8a).

into these categories. This is in contrast to the subjects' data where only 12% of the fixations were classified into those five categories. The feature model does no better at predicting the distribution of actual fixations. The most notable characteristic of the feature model distribution (Fig. 8c) is the large number of fixations in the "ahead" category (32%), an area that happens to contain a door. In practice, only 5% of the subjects' fixations were classified in the "ahead" category. In addition to doors, the feature model predicted a number of fixations on posters, presumably due to the fact that poster edges have features similar to the target, i.e., vertical and large. Fig. 8d shows the geographic model distribution. The truncated shape reveals the left-side restriction and resembles, in part, the skewed distribution of the actual fixations. But unlike the actual fixation distribution where fixations are primarily on the doors, the majority of the fixations predicted by the geographic model are classified in the categories "ahead", "ceiling", and the left-side posters/walls. The distribution for the feature–geographic model (Fig. 8e) has the truncated shape of the geographic model but a lower number of fixations in the "ahead" and "ceiling" categories, due to the bias for "vertical" and "large".

A hierarchical clustering method was used to determine whether there was an obvious pattern in subjects' fixation behavior among the various categories (e.g. left doors, right doors, floor). Hierarchical clustering is a multivariate technique that groups together elements that have similar values. In our case the categories with similar fixation percentages were grouped together. The process starts with each element as its own cluster and the distance between each cluster is calculated. The two clusters that are closest together are combined and the process reiterates until all points are in a final cluster. The clustering tree can be cut at various points. A cut at two produces a cluster consisting of the left-side doors and a cluster consisting of the other categories. With three clusters, the cluster of left-side doors breaks down into two clusters: the near left-side doors (doorL1, doorL2, and doorL3) and the far left-side doors (doorL4 and doorL5). A cut at four produces the clustering represented in Fig. 9. The clusters are differentiated by different symbols (and each datum is labeled with its specific category code). Listed in order of percentage from most fixated to least, the four clusters consist of the following: (1) the far left-side doors, (2) the near left-side doors, (3) left-side posters and walls, nearest right-side

door, ahead, and floor, and (4) the right-side doors, right-side posters and walls, and ceiling.

Fig. 9 shows how the predicted fixation distribution of the various models relates to the actual fixation distribution. Fig. 9a plots the distribution of the visual salience model against the actual fixation distribution. The correlation coefficient, r, for the two distributions was −0.23, ns, indicating no significant linear relationship between the predictions of the visual salience model and where the subjects actually looked while walking. The bivariate normal density ellipse shows where 95% of the data are expected to lie. The one category that falls outside the density contour is the target. The target is fixated more frequently than the model predicts. No significant linear relationship was found between the actual fixation distribution and the feature model distribution, $r = -0.12$, ns (Fig. 9b) or the geographic model distribution, $r = 0.19$, ns (Fig. 9c). The only significant linear relationship that was found between the actual fixation distribution and a model prediction distribution was for the feature–geographic model, $r = 0.53$, $p < 0.05$, (Fig. 9d). The bivariate fit shows that two categories fall outside the 95% density contour, the nearest wall/poster on the left side (wallL1) and the target. With this model, the nearest wall/poster on the left side is fixated much less frequently than the model predicts. And the target is fixated much more frequently than the model predicts; this was true for all the models.

4. Discussion

In this study, we evaluated, against a baseline case of random scanning, how well various oculomotor strategies predict the gaze patterns of subjects while walking. Eye and scene images were recorded as each subject walked to a pre-defined target. From these recordings, the direction of gaze (eye-on-scene) was determined and fixations were identified. For each fixation, the direction of gaze was classified into one of 20 categories, producing a sequence of direction-of-gaze categories represented by a string of letters. Each model generated predicted fixation locations that were classified in the same manner as were the actual fixations. An optimal alignment was determined for each subject's gaze pattern and each model's output, and the percent similarity between the data and model was calculated from the number of matched pairs.

Both versions of random scanning—one in which the random scanner chose x and y fixation coordinates from anywhere on the image and another in which the distance between fixations was randomly selected from a distribution of real eye-movement amplitudes—matched about a quarter (22%–23%) of the gaze patterns of the subjects.

It is unclear why the "realistic" version of the random-scanner did not outperform the "totally random" version in matching the subjects' gaze patterns. The random scanner in the "realistic" version had knowledge of the real eye-movement amplitudes, and fixations were drawn from that distribution. The fact that the random scanner's selection of fixation *direction* was unrestricted must have overwhelmed any advantage from the knowledge of amplitude. Had we analyzed the degree of similarity between model and data on a pixel basis rather than category basis we might have found a superiority effect of the "realistic" version compared to the "totally random" version.

The visual salience model matched about the same percentage of the subjects' gaze patterns as did the random scanner, 21.3%. The comparability in predictive power of the visual salience model and the random scanner demonstrates that an oculomotor strategy based on the visual salience of the image is no better at predicting human fixation behavior in this task than an oculomotor strategy that randomly selects image locations.

The feature model, also, performed at the same level as the random scanner in predicting the subjects' gaze patterns. Only 20.8% of the feature model's predictions matched the gaze patterns of the subjects. The low performance of the feature model may be due in part from our selection of target features, vertical and large. These coarsely defined features did not uniquely specify the target. The model often chose the posters' edges as the fixation location.

The geographic model outperformed the random scanning models, matching about a third of the subjects' gaze patterns. This improvement in predictive power suggests that the subjects used the information about the general vicinity of a target to guide their fixations. However, that fixations sometimes fell on the right side of the image indicates that the subjects did not feel constrained to fixate *only* the left side. To implement the geographic model, we adopted a liberal interpretation of the terms "left side". Any spatial location on the left side *of the image* qualified as an acceptable fixation location. This interpretation was perhaps too coarse for the current task since in real life left-side doors are located in the walls. The too-coarse interpretation could explain the higher percentage of fixations in the categories: "ahead", "ceiling", and "floor" (35%) for the geographic model compared to the subjects' actual fixations (9%).

The feature–geographic model best predicted the subjects' gaze patterns. With this model, the image features common to the target, vertical and large, were heavily weighted, and gaze was restricted to the left side of the image. Even though both sets of information were coarsely defined, together they were sufficient to increase the predictive power of the model to nearly twice the level of a random scanner, visual salience model, or

feature model. The results showed that the feature–geographic model predicted nearly half (47.5%) the gaze patterns of the subjects. Furthermore, a linear relationship between the distributions of fixation percentages for the feature–geographic model and the actual fixations was demonstrated by a significant correlation ($r = 0.53$). This is remarkable given the lax constraint of the current task and the coarse feature and geographic information. Presumably the predictive power of this model would increase even more with a more refined description of the target's features. But, even with an improved description of the target's features, it is unlikely that the feature–geographic model would be able to fully predict the subjects' gaze patterns in the current task. The model lacks a specified "procedure" that the subjects may have used to carry out the task (Suppes, 1990). An inspection of the subjects' gaze patterns reveals that a significant number of fixations were on previously fixated doors (62%). The task of finding the fifth door on the left requires counting the doors and maintaining in memory the count. The behavior of "looking back" may be related to rehearsing or refreshing one's memory. This idea receives some support from other eye-movement studies in which working memory was required and re-fixation behavior was seen (Ballard, Hayhoe, & Pelz, 1995; Land et al., 1999; Land & Lee, 1994). Thus, a model that uses information about a target's features and its general vicinity does a fairly good job predicting fixation behavior, but procedural knowledge may be required to more fully capture the gaze patterns of subjects' performing an everyday activity.

Walking to a target was chosen as the real-world task, based on the desire to balance the attention demand of the appointed task. Walking down a hallway toward a goal is not as demanding as driving around a bend in the road or making tea, yet it does require more attention than free viewing. In our study, the target had to be detected among similar and dissimilar distractors and the subject had to walk to it. In this relatively unconstrained situation the subject had sufficient time to look around. In practice, the subjects primarily fixated the left side of the image (see Fig. 7). However, in a few instances the subjects fixated elsewhere (total of 34). One might expect that if visual salience played a role in directing fixations in the present task, in the instances where the subjects did not fixate on the left side gaze would have been guided by the visual salience of the image. This was not the case. In only 7 of the 34 instances did the classification of the visual salience model prediction match the category of the actual fixation location. This finding suggests that visual salience, alone, is not a very useful concept in the present study.

The oculomotor strategy that one uses may depend on the rigor of the task demands (attention and time) and on the ease in detecting the task-relevant components. An observation by Land and Lee (1994) lends some support to this idea. When the task demand of driving was high (around a bend in the road) drivers' fixations were tightly bound to task-relevant information but when the demand was relaxed (wide road driving) one driver had many fixations on driving-irrelevant information. While it is beyond the scope of this paper, a more comprehensive assessment of the oculomotor strategies in real-world tasks would include various oculomotor strategies tested against the gaze patterns of subjects performing a broad cross-section of tasks.

4.1. Relation to other studies

The question of whether visual search theories based on findings of laboratory-based experiments have any application to the real world has been addressed previously. Wolfe (1994) explored the issue in a manner very different from the present study. He expanded the type of visual stimuli in his visual search experiments to include more "naturalistic" stimuli. The stimuli were computer-generated graphics that resembled aerial views of terrain (e.g. rivers, lakes). The subjects' task was the same as in the traditional visual search experiments—to find a target embedded in a background. Although Wolfe's study has more differences than similarities to the present study, the rationale for both studies was the same, to examine the generalizability of the laboratory-based visual search strategies. Wolfe's conclusion was that the rules of visual search defined by artificial stimuli in laboratory experiments do apply to the continuous, naturalistic stimuli and may extend to more real-world situations.

Parkhurst et al. used the visual salience model that we tested in the present study but they arrived at a different conclusion concerning the contribution of the visual salience model (Parkhurst et al., 2002). In their study, subjects freely viewed images of natural and artificial scenes while their eye movements were recorded. The stimulus salience at fixation locations was computed and compared to the mean salience expected by chance (computed from saliency values randomly chosen from the saliency map). The average salience computed from the fixation locations was higher than that expected by chance alone. Parkhurst et al. interpreted the results as providing evidence that stimulus-driven, bottom-up mechanisms contribute significantly to guiding attention in natural viewing.

The interpretation of Parkhurst et al. is at odds with our finding that the visual salience model performed *no better* than expected by chance (i.e., the random scanner). However, several differences exist between the two studies, and it could be that one or more is responsible for the apparent discrepancy. One difference between the

studies is the type of analysis performed. In our study we were interested in knowing how well the visual salience model predicted where in the scene people directed their fixation while walking toward a pre-defined target. To this end, we compared the scene categories of the actual fixation locations to the scene categories of the predicted fixation locations of the visual salience model. Parkhurst et al. used a converse approach in assessing the role of salience in directing gaze. Rather than comparing the location of the highest salience in each image to the actual fixation location, they determined the salience at the fixation location and compared its magnitude to the average salience across randomly chosen locations in the saliency map. This is not the same comparison as performed in the present study nor is the information it provides the same.

Another difference between studies is the task given to the subjects. In the Parkhurst et al. study, subjects were instructed to "look around" at an image. The stimulus was a static image that was freely viewed for a period of 5 s. The authors claimed that by avoiding specific instructions to the subjects they would more likely do what they normally do when looking at images. However, one might argue that without having a specific task in mind (unlike what occurs in everyday experience) subjects resorted to using the only thing available to them to guide their fixation, i.e., image salience. In our study, subjects walked toward a pre-defined target and the stimulus was a sequence of continuously changing images. The task was very specific and the target well defined.

4.2. Limitations of the study

The mobility route that we chose for this study was very simple. Apart from the subject, the environment contained no moving objects (e.g. people, cars) or abrupt fluctuations in environmental conditions (e.g. drop-offs, severe illumination changes). While this choice was deliberate in an attempt to minimize the number of variables in the study, the simplicity of the route may be viewed as a limitation. This simple route does not represent the range of environments that we typically encounter. Different environmental conditions may produce different gaze patterns. For example, the introduction of new objects into the scene may grab attention and re-direct gaze. Laboratory studies have shown that newly appearing objects are powerful attractants for attention and fixations (Yantis & Hillstrom, 1994; Yantis & Jonides, 1984, 1996).

The subjects in our study were moving observers, which raises a potential technicality in the way we tested the models. The movement of an observer produces optic flow—a change in the pattern of light intensities reflected from objects in the environment to the observer's eye. This motion is a real input to the subject's visual system that we did not include as input to the models. The models were fed static video images—frames of each subject's scene movie. (Note that this input actually favors the models since the image is already restricted to that selected by the subject via head movements.) If the motion in optic flow patterns plays a key role in directing gaze our test would have been biased since we would have omitted an essential component of the stimulus. However, in fairness to our approach, none of the models that we tested were designed to take into account the motion feature. If optic flow does play a role in directing gaze it is unclear what aspect of the pattern is used for fixation (see Cutting, 1986 for suggestions). Moreover at the slow travel speeds of our subjects (range of 0.8–1.3 m/s) it is unclear whether the velocity vectors generated from fixating anything other than nearby objects would be useful.

In conclusion, an oculomotor search strategy that allows for "top-down" guidance from coarse geographic and featural information better predicts the visual scanning behavior of subjects walking toward a target compared to a random scanning strategy or one based solely on "bottom-up" stimulus driven factors.

Acknowledgements

This research was supported by the National Institutes of Health, National Eye Institute, under grant EY07839 to KAT. The authors thank Lauren Itti for helpful instructions and use of his computer implementation of the visual salience model. We also thank Julie Stahl for her role in data collection and analysis, Marc Shapiro and Claudius Li for software development, and Frank Turano for sequence analyses.

References

Adelson, E. H., Anderson, C. H., Bergen, J. R., Burt, P. J., & Ogden, J. M. (1984). Pyramidal methods in image processing. *RCA Engineer, 29*(6), 33–41.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representation in natural tasks. *Journal of Cognitive Neuroscience, 7*, 66–80.

Burt, P. J., & Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications COM, 31*(4), 532–540.

Buswell, G. T. (1935). *How people look at pictures: a study of the psychology of perception in art*. Chicago, IL: The University of Chicago Press.

Choi, Y. S., Mosley, A. D., & Stark, L. W. (1995). String editing analysis of human visual search. *Optometry and Vision Science, 72*, 439–451.

Corbetta, M. (1998). Frontoparietal cortical networks for directing attention and the eye to visual locations: identical, independent, or overlapping neural systems? *Proceedings of the National Academy of Sciences of the United States of America, 95*, 831–838.

Cutting, J. E. (1986). *Perception with an eye for motion*. Cambridge, Mass: The MIT Press.

Epelboim, J., Steinman, R. M., Kowler, E., Pizlo, Z., Erklens, C. J., & Collewijn, H. (1997). Gaze-shift dynamics in two kinds of sequential looking tasks. *Vision Research, 37*, 2597–2607.

Ferris, F. L., Kassoff, A., Bresnick, G., & Bailey, I. (1982). New visual acuity charts for clinical research. *American Journal of Ophthalmology, 94*, 91–96.

Geisler, W. S., & Chou, K. (1995). Separation of low-level and high-level factors in complex tasks: visual search. *Psychological Review, 102*, 356–378.

Hayhoe, M. M., Bensinger, D. G., & Ballard, D. H. (1998). Task constraints in visual working memory. *Vision Research, 38*, 125–137.

Hoffman, J. E. (1978). Search through a sequentially presented visual display. *Perception and Psychophysics, 23*, 1–11.

Hoffman, J. E. (1979). A two-stage model of visual search. *Perception and Psychophysics, 25*, 319–327.

Inhoff, A., Pollatsek, A., Posner, M., & Rayner, K. (1989). Covert attention and eye movements during reading. *Quarterly Journal of Experimental Psychology A, 41*, 63–89.

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

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 (PAMI), 20*, 1254–1259.

Kito, T., Haraguchi, M., Funatsu, T., Sato, M., & Kondo, M. (1989). Measurements of gaze movements while driving. *Perceptual and Motor Skills, 68*, 19–25.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology, 4*, 219–227.

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research, 35*, 1897–1916.

Kundel, H. L., Nodine, C. F., Thickman, D., & Toto, L. (1987). Searching for lung nodules. A comparison of human performance with random and systematic scanning models. *Investigative Radiology, 22*, 417–422.

Kustov, A., & Robinson, D. (1996). Shared neural control of attentional shifts and eye movements. *Nature, 384*, 74–77.

Land, M., & Horwood, J. (1995). Which parts of the road guide steering? *Nature, 377*, 339–340.

Land, M., & McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience, 3*, 1340–1345.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception, 28*, 1311–1328.

Land, M. F. (1992). Predictable eye–head coordination during driving. *Nature, 359*, 318–320.

Land, M. F. (1998). The visual control of steering. In L. R. A. J. Harris & A. Micheal (Eds.), *Vision and action*. Cambridge: Cambridge University Press.

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature, 369*, 742–744.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Niebur, E., & Koch, C. (1996). Control of selective visual attention: modeling the where pathway. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems* (vol. 8, pp. 802–808). Cambridge, MA: MIT Press.

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

Pelli, D. G., Robson, J. G., & Wilkens, A. J. (1988). The design of a new letter chart for measuring contrast sensitivity. *Clinical Vision Sciences, 2*, 187–199.

Posner, M., & Petersen, S. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*, 25–42.

Shepherd, M., Findlay, J. M., & Hockey, R. J. (1986). The relationship between eye movements and attention. *Quarterly Journal of Experimental Psychology, 38A*, 475–491.

Stark, L., & Choi, Y. S. (1996). Experimental metaphysics: the scanpath as an epistemological mechanism. *Visual Attention and Cognition*.

Suppes, P. (1990). Eye-movement models for arithmetic and reading performance. In E. Kowler (Ed.), *Reviews of oculomotor research: eye movements and their role in visual and cognitive processes* (vol. 4, pp. 455–478). New York: Elsevier.

Theeuwes, J., & Burger, R. (1998). Attentional control during visual search: the effect of irrelevant singletons. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 1342–1353.

Toet, A., Kooi, F. L., Bijl, P., & Valeton, J. M. (1998). Visual conspicuity determines human target acquisition performance. *Optical Engineering, 37*, 1969–1975.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97–136.

Wann, J., & Swapp, D. (2000). Why you should look where you are going. *Nature Neuroscience, 3*, 647–648.

Wolfe, J. M. (1994). Visual search in continuous, naturalistic simuli. *Vision Research, 34*, 1187–1195.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 419–433.

Yantis, S., & Hillstrom, A. P. (1994). Stimulus-driven attentional capture: evidence from equiluminant visual objects. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 95–107.

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 601–621.

Yantis, S., & Jonides, J. (1996). Attentional capture by abrupt onsets: new perceptual objects or visual masking? *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1505–1513.

Zelinsky, G. J., & Sheinberg, D. L. (1997). Eye movements during parallel–serial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 244–262.