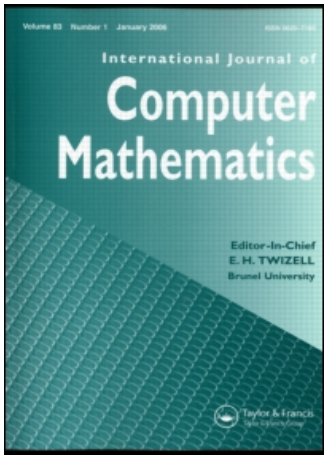

International Journal of Computer Mathematics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713455451

Perceptual image retrieval using eye movements

Online Publication Date: 01 September 2007
To cite this Article: Oyekoya, Oyewole and Stentiford, Fred (2007) 'Perceptual image retrieval using eye movements', International Journal of Computer Mathematics, 84:9, 1379 - 1391
To link to this article: DOI: 10.1080/00207160701242268
URL: http://dx.doi.org/10.1080/00207160701242268

International Journal of Computer Mathematics
Vol. 84, No. 9, September 2007, 1379–1391

Perceptual image retrieval using eye movements

OYEWOLE OYEKOYA* and FRED STENTIFORD

University College London, Adastral Park, Ipswich IP5 3RE, UK

(Received 16 September 2006; revised version received 11 December 2006; accepted 22 January 2007)

This paper explores the feasibility of using an eye tracker as an image retrieval interface. A database of 1000 Corel images is used in the study and results are analysed using ANOVA. Results from participants performing image search tasks show that eye tracking data can be used to reach target images in fewer steps than by random selection. The effects of the intrinsic difficulty of finding images and the time allowed for successive selections were also considered. The results indicated evidence of the use of pre-attentive vision during visual search.

Keywords: Computer interface; Image retrieval; Visual attention; Eye movements; Search performance

AMS Subject Classifications: 68U10; 62H35; 62J10

1. Introduction

Images play an increasingly important part in the lives of many people. There is a critical need for automated management, as the flow of digital visual data increases and is transmitted over the network. Retrieval mechanisms must be capable of handling the amount of data efficiently and quickly. Existing systems are capable of retrieving archiving material according to date, time, location, format, file size, etc. However, the ability to retrieve images with semantically similar content from a database is more difficult.

One of the major issues in information searching is the problem associated with initiating a query. The lack of high-quality interfaces for query formulation is a further barrier to effective image retrieval systems [1]. Eye tracking presents an adaptive approach that can capture the user's current needs and tailor the retrieval accordingly. Understanding the movement of the eye over images is an essential component in this research.

Research in the applications of eye tracking is increasing, as presented in Duchowski's review [2] of diagnostic and interactive applications based on offline and real-time analysis, respectively. Interactive applications have concentrated upon replacing and extending existing computer interface mechanisms rather than creating a new form of interaction. The tracking of eye movements has been employed as a pointer and a replacement for a mouse [3], to vary the screen scrolling speed [4] and to assist disabled users [5]. Dasher [6] uses a method for text entry that relies purely on gaze direction. In its diagnostic capabilities, eye tracking provides

*Corresponding author. Email: o.oyekoya@adastral.ucl.ac.uk

International Journal of Computer Mathematics
ISSN 0020-7160 print/ISSN 1029-0265 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/00207160701242268

a comprehensive approach to studying interaction processes such as the placement of menus within web sites and to influence design guidelines more widely [7]. The imprecise nature of saccades and fixation points has prevented these approaches from yielding benefits over conventional human interfaces. Fixations and saccades are used to analyse eye movements, but it is evident that the statistical approaches to interpretation (such as clustering, summation and differentiation) are insufficient for identifying interests due to the differences in people's perception of image content [8]. There has been some recent work on document retrieval in which eye tracking data has been used to refine the accuracy of relevance predictions [9]. Applying eye tracking to image retrieval requires that new strategies be devised that can use visual and algorithmic data to obtain natural and rapid retrieval of images.

Traditional approaches of image retrieval suffer from three main disadvantages. Firstly, there is a real danger that the use of any form of predefined feature measurements will be unable to handle unseen material. Image retrieval systems normally rank the relevance between a query image and target images according to a similarity measure based on a set of features. Pre-determined features can take the form of edges, colour, location, texture, and others. Secondly, the choice of low-level features is unable to anticipate a user's high-level perception of image content. This information cannot be obtained by training on typical users because every user possesses a subtly different subjective perception of the world and it is not possible to capture this in a single fixed set of features and associated representations. Thirdly, descriptive text does not reflect the capabilities of the human visual memory and does not satisfy users' expectations. Furthermore, the user may change his/her mind and may also be influenced by external factors. An approach to visual search should be consistent with the known attributes of the human visual system and account should be taken of the perceptual importance of visual material. Recent research in human perception of image content [10] suggests the importance of semantic cues for efficient retrieval. Relevance feedback mechanisms [11] is often proposed as a technique for overcoming many of the problems faced by fully automatic systems by allowing the user to interact with the computer to improve retrieval performance. This reduces the burden on unskilled users to set quantitative pictorial search parameters or to select images (using a mouse) that come closest to meeting their goals. This has prompted research into the viability of eye tracking as a natural input for an image retrieval system. Visual data can be used as input as well as a source of relevance feedback for the interface. Human gaze behaviour may serve as a new source of information that can guide image search and retrieval. Human eye behaviour is defined by the circumstances in which they arise. The eye is attracted to regions of the scene that convey what is thought at the time to be the most important information for scene interpretation. Initially, these regions are pre-attentive in that no recognition takes place, but moments later in the gaze the fixation points depend more upon either our own personal interests and experience or a set task. Humans perceive visual scenes differently. We are presented with visual information when we open our eyes and carry out non-stop interpretation without difficulty. Research in the extraction of information from visual scenes has been explored by Yarbus [12], Mackworth and Morandi [13] and Hendersen and Hollingworth [8]. Mackworth and Morandi [13] found that fixation density was related to the measure of informativeness for different regions of a picture and that few fixations were made to regions rated as uninformative. The picture was segmented and a separate group of observers were asked to grade the rate of informativeness. Scoring the informativeness of a region provides a good insight into how humans perceive a scene or image. Henderson and Hollingworth [8] described semantic informativeness as the meaning of an image region and visual informativeness as the structural information. Fixation positions were more influenced by the former compared with the latter. The determination of informativeness and corresponding eye movements is influenced by task demands [12].

Previous work [14] used a visual attention model to score the level of informativeness in images and found that a substantial part of the gaze of the participants during the first two seconds of exposure is directed at informative areas as estimated by the model. This lent credence to the belief that the gaze information obtained from users when presented with a set of images could be useful in driving an image retrieval interface. More recent work [15] compared the performance of the eye and the mouse as a source of visual input. Results showed faster target identification for the eye interface than the mouse for identifying a target image on a display.

In this paper, experiments are described that explore the viability of using the eye to drive an image retrieval interface. Preliminary work was reported in [16]. In a visual search task users are asked to find a target image in a database and the number of steps to the target image are counted.

2. Methodology

In this system eye movement is used to formulate queries for Content Based Image Retrieval (CBIR) processing. It is intended that this should provide a rapid and natural interface for searching visual digital data in an image database. A network of links (see figure 3) between images in an image collection is computed by calculating a similarity measure between all possible pairs of images. The network of links possessing the highest values may be traversed very rapidly using eye tracking providing the users' gaze behaviours yield suitable information about their intentions. Users will tend to look at the objects in which they are interested during a search and this provides the machine with the necessary information to retrieve plausible candidate images by referring to the pre-computed similarity link values. Retrieved images will contain regions that possess similarity links with the previously gazed regions and can be presented to the user in a variety of ways. The user might navigate the similarity links as illustrated by the green arrow/path in figure 3.

2.1 *Data and equipment*

One thousand images were selected from the Corel image library. Images of 127 kilobytes and 256×170 pixel size were loaded into the database. The categories included boats, landscapes, vehicles, aircraft, birds, animals, buildings, athletes, people and flowers. The initial screen (including the position of the target image) is shown in figure 1. Images were displayed as 229×155 pixel size in the 4×4 grid display.

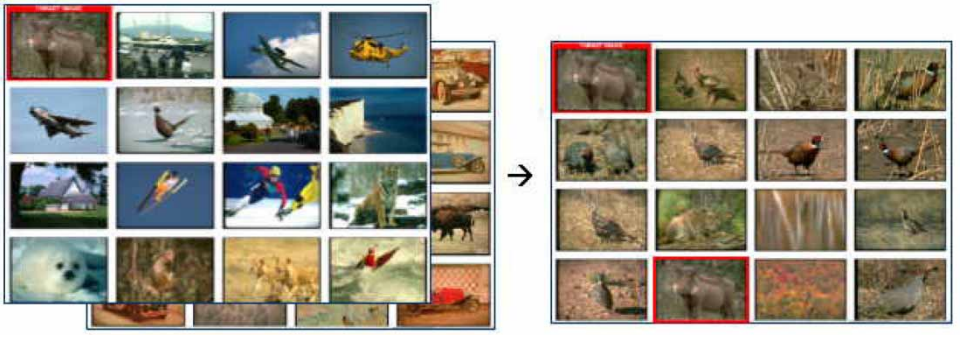

Figure 1. Initial screen leading to the final screen with retrieved target.

An Eyegaze System [17] was used in the experiments to generate raw gaze point location data at a camera field rate of 50 Hz (units of 20 ms). A clamp with a chin rest provided support for the chin and forehead in order to minimize the effects of head movements, although the eye tracker does accommodate head movement of up to 1.5 inches (3.8 cm). Calibration is needed to measure the properties of each subject's eye before the start of the experiments. The images were displayed on a 15 inch LCD Flat Panel Monitor at a resolution of 1024 × 768 pixels. The loading of 16 images in the 4 × 4 grid display took an average of 100 ms on a Pentium IV 2.4 GHz PC with 512 MB of RAM. Gaze data collection and measurement of variables were suspended while the system loaded the next display. The processing of information from the eye tracker is done on a 128 MB Intel Pentium III system with a video frame grabber board.

Images are presented in a 4 × 4 grid with target image presented in the top left corner of the display. The user is asked to search for the target image and on the basis of the captured gaze behaviour the machine selects the most favoured image. The next set of 15 images is then retrieved from the database on the basis of similarity scores and displayed for the next selection. The session stops when the target image is found or a prescribed number of displays is reached. A similarity measure [18] was used to pre-compute a network of similarity scores between all pairs of images in the database.

2.2 *Similarity measure*

Image retrieval systems normally rank the relevance between a query image and target images according to a similarity measure based on a set of features. The similarity measure [18] used in this work, termed Cognitive Visual Attention (CVA model), is not dependent upon intuitively selected features, but instead upon the notion that the similarity of two patterns is determined by the number of features in common. This means that the measure can make use of a virtually unlimited universe of features rather than a tiny manually selected subset that will be unable to characterize many unseen classes of images. Moreover, the features are deliberately selected from image regions that are salient according to the model and, if validated, reflect similarity as judged by a human. The CVA model relies upon the matching of large numbers of pairs of pixel groups (called *forks* here) taken from patterns A and B under comparison (figure 2).

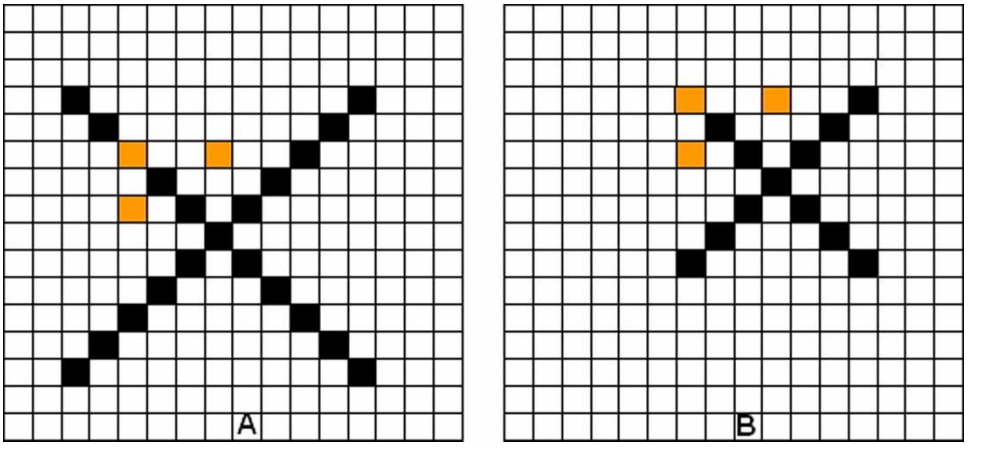

Figure 2. Neighbourhood at location x mismatching at y.

Let a pixel $\boldsymbol{x}$ in a pattern correspond to colour components $\boldsymbol{a}$

$$\boldsymbol{x} = (x_1, x_2), \quad \text{where } \boldsymbol{a} = (a_1, a_2, a_3).$$

Let $\boldsymbol{F}(x) = \boldsymbol{a}$. Select a fork of m random points S_A in pattern A (e.g. the three pixels shown in figure 2) where

$$S_\mathrm{A} = \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_m.$$

Likewise, select a fork of m points S_B in pattern B where

$$S_\mathrm{B} = \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_m$$

and

$$\boldsymbol{x}_i - \boldsymbol{y}_i = \boldsymbol{\delta}.$$

S_B is a translated version of S_A. The fork S_A matches fork S_B if

$$|F_j(\boldsymbol{x}_i) - F_j(\boldsymbol{y}_i)| < \varepsilon_j, \quad \forall\, i \text{ for some displacement } \boldsymbol{\delta}.$$

In general, $\boldsymbol{\epsilon}$ is not a constant and will be dependent upon the measurements under comparison

$$\epsilon_j = f_j(\boldsymbol{F}(\boldsymbol{x}), \boldsymbol{F}(\boldsymbol{y})).$$

In addition, it is required that $|F_k(\boldsymbol{x}_i) - F_k(\boldsymbol{x}_j)| > \epsilon_k$ for some k, $i \neq j$ so that some pixels in S_A mismatch each other and the similarity measure is taken over regions of high attention and not just on areas of sky, for example.

In effect, up to N selections of the displacements $\boldsymbol{\delta}$ apply translations to S_A to seek a matching fork S_B. The CVA similarity score C_AB is produced after generating and applying T forks S_A:

$$C_\mathrm{AB} = \sum_{i=1}^{T} w_i,$$

where $w_i = 1$ if S_A matches fork S_B or 0 otherwise.

C_AB is large when a large number of forks are found to match both patterns A and B and represents features that both patterns share. In other words, the CVA similarity score is incremented each time one of the set of pixel sets matches a set in pattern B. This means that image pairs A,B which possess large numbers of matching forks will obtain high CVA scores by virtue of the number of such features they possess in common. It is important to note that if C_AC also has a high value it does not necessarily follow that C_BC is large because patterns B and C may still have no features in common. The measure is not constrained by the triangle inequality. The CVA algorithm was applied to the 1000 images to pre-compute similarity scores for all pairs of images to obtain a network of similarity links (figure 3).

2.3 *Random selection strategy*

A random selection strategy was employed to provide a performance baseline which a more intelligent approach would need to exceed. The automatic random selection tool randomly selected an image from each successive screen holding 15 displayed images rather than by eye gaze. The random selection tool was applied with each of the 1000 images acting as the target image and the number of steps to target recorded. It was found that, as the number of randomly retrieved images in each display was increased, the likelihood of finding the target image also increased.

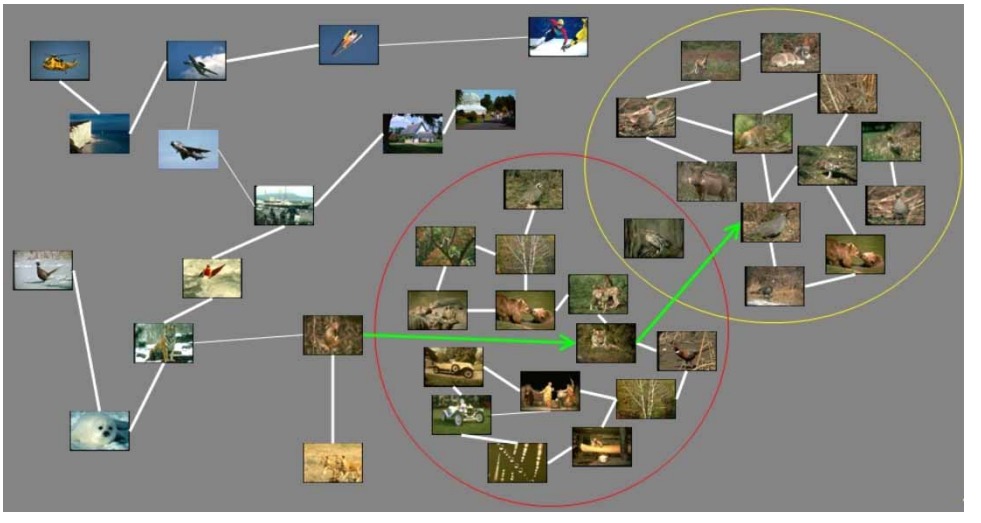

Figure 3. Representation of pre-computed similarity links where strength is indicated by line thickness.

Two strategies were employed to assist in the selection of target images of varying difficulty for search experiments. Firstly, a plot of the frequency distribution of steps to target for every image in the database revealed those images that were frequently found in the fewest and most number of steps. Secondly, a plot of the frequency distribution of the 15 images with the highest similarity scores with each image in the database indicated those images that were similar to most other images and were therefore most likely to be found when traversing similarity links during a search. By analysing the search performance and the retrieved image sets, the two strategies revealed the easy-to-find and hard-to-find images.

3. Experiment 1

3.1 *Selection of target images*

Four of the easy-to-find images and four of the hard-to-find images were picked as target images for this experiment. These are shown in figure 4.

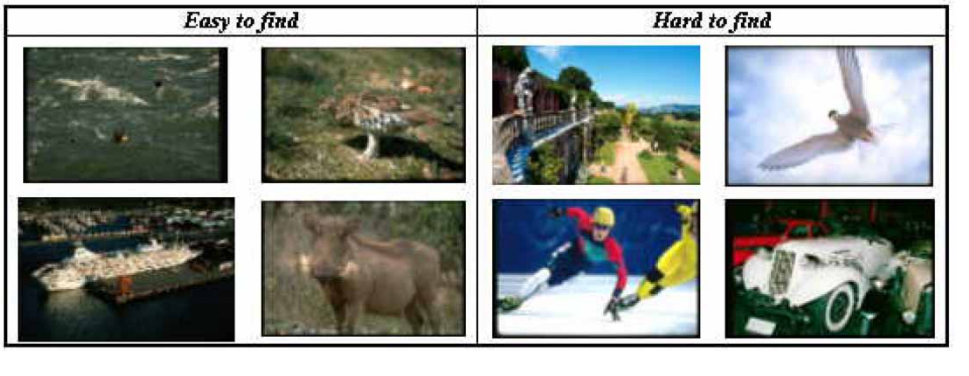

Figure 4. Target images.

3.2 *Experiment design*

Thirteen unpaid participants took part in the experiment. Participants included a mix of students and university staff. All participants had normal or corrected-to-normal vision and provided no evidence of colour blindness.

One practice run was allowed to enable better understanding of the task and to equalize skill levels during the experiment. Participants understood that there would be a continuous change of display until they found the target, but did not know what determined the display change. The display change is triggered by the sum of all fixations of 80 ms and above on an image position exceeding a fixation threshold. Two fixation thresholds of 400 ms and 800 ms were employed as a factor in the experiment. The display included either no randomly retrieved image (all 15 images are selected on the basis of similarity scores) or one randomly retrieved image (one image is randomly selected from the database). Participants performed eight runs, using both image types (easy-to-find and hard-to-find). Four treatment combinations of the two fixation thresholds (400 ms and 800 ms) and two randomly retrieved levels (0 and 1) were applied to each image type. Any sequence effect was minimized by randomly allocating each participant to different sequences of target images. The first four runs were assigned to each image type. There was a 1 min rest between runs. The maximum number of steps to target was limited to 26 runs.

3.3 *Results and analysis*

Three dependent variables, the number of steps to target, the time to target (F_1), and the number of fixations (F_2) of 80 ms and above, were monitored and recorded during the experiment. Eight dependent variables were recorded for each participant. The average figures are presented in table 1.

One hundred and four figures were entered for each dependent variable into repeated measures ANOVA with three factors (image type, fixation threshold and randomly retrieved).

The results of the ANOVA performed on the steps to target revealed a significant main effect of image type, $F(1, 12) = 23.90$, $p < 0.0004$, with fewer steps to target for easy-to-find images (14 steps) than the hard-to-find images (22 steps). Easy-to-find target images were found in fewer steps by participants than the hard-to-find images as predicted by the evidence obtained using the random selection strategy.

The main effect of the fixation threshold was not significant with $F(1, 12) = 1.50$, $p < 0.25$. The main effect of randomly retrieved was also not significant, $F(1, 12) = 0.17$, $p < 0.69$. Generally, the influence of including one randomly retrieved image in each display produced

Table 1. Analysis of human eye behaviour on the interface (rounded-off mean figures).

Image type	Fixation threshold	Randomly retrieved	A	B	C	D
Easy-to-find	400 ms	0	38.5%	14	34.9	99
		1	53.8%	18	36.8	109
	800 ms	0	38.5%	14	55.8	153
		1	15.4%	11	51.3	140
Hard-to-find	400 ms	0	69.2%	23	52.7	166
		1	84.6%	23	50.0	167
	800 ms	0	92.3%	24	105.0	327
		1	69.2%	19	83.5	258

A = target not found (frequency); B = steps to target; C = time to target; D = fixation numbers.

little or no difference in the steps to target, time to target and fixation numbers. Even when compared with the random selection tool, the steps to target did not significantly differ. All two-factor and three-factor interactions were not significant.

The analysis of the time to target produced similar results to the analysis of the number of fixations. There was a significant main effect of image type, $F_1(1, 12) = 24.11$, $p < 0.0004$, $F_2(1, 12) = 21.93$, $p < 0.0005$, with shorter time to target and fewer fixations for easy-to-find images (40.5 s and 125 fixations) than the hard-to-find images (71.3 s and 229 fixations). The main effect of the fixation threshold was also similarly significant with $F_1(1, 12) = 18.27$, $p < 0.001$ and $F_2(1, 12) = 16.09$, $p < 0.002$. There were more fixations and more time was spent on hard-to-find images than the easy-to-find images. This is consistent with the conclusion of Fitts *et al.* [19] that complex information leads to longer fixation durations and higher fixation numbers.

In line with the steps to target, the main effect of randomly retrieved was also not significant, $F_1(1, 12) = 1.49$, $p < 0.25$ and $F_2(1, 12) = 0.76$, $p < 0.40$.

Image type interacted with the fixation threshold, $F_1(1, 12) = 8.04$, $p < 0.015$ and $F_2(1, 12) = 5.84$, $p < 0.032$, and an analysis of simple main effects indicated a significant difference in time to target and fixation numbers for the fixation thresholds when hard-to-find images were presented, $F_1(1, 12) = 20.00$, $p < 0.001$ and $F_2(1, 12) = 16.25$, $p < 0.002$, but, interestingly, no significant difference when easy-to-find images were presented, $F_1(1, 12) = 3.62$, $p < 0.08$ and $F_2(1, 12) = 3.57$, $p < 0.08$. There was no significant difference in the time to target and fixation numbers between the threshold levels for the easy-to-find images as opposed to the hard-to-find images. In other words, setting a higher threshold did not significantly differ when either 400 ms or 800 ms was used for the easy-to-find images, but it did for the hard-to-find images. However, the steps to target did differ for both image types under either of the threshold conditions. A future experiment will be needed to investigate whether the thresholds can be reduced further, at least for the easy-to-find images.

The same treatment combinations experienced by all participants were applied to the random selection tool to obtain 104 dependent variables (steps to target). By combining the variables, 208 figures were entered into a mixed design multivariate ANOVA with two observations per cell and three factors (selection mode, image type and randomly retrieved). The average figures are presented in table 2.

In summary, the results of the ANOVA revealed a main effect of the selection mode, $F(2, 23) = 3.81$, $p < 0.037$, with fewer steps to target when the eye gaze is used (18 steps) than when random selection is used (22 steps). There was also a main effect of image type, $F(2, 23) = 28.95$, $p < 0.00001$, with fewer steps to target for easy-to-find images (16 steps) than the hard-to-find images (24 steps).

Table 2. Comparison of eye and random selection (rounded-off mean figures).

Selection mode	Image type	Randomly retrieved	A	B
Eye gaze	Easy-to-find	0	38.5%	14
		1	34.6%	15
	Hard-to-find	0	80.8%	23
		1	76.9%	21
Random selection	Easy-to-find	0	57.7%	20
		1	38.5%	16
	Hard-to-find	0	96.2%	25
		1	92.3%	26

A = target not found (frequency); B = steps to target.

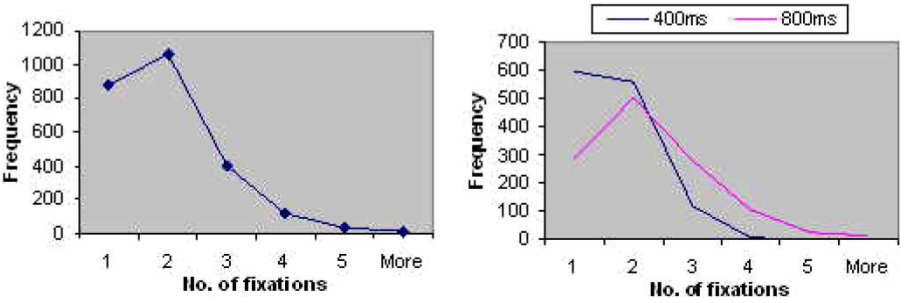

Figure 5. Histogram plots of the number of fixations on selected images.

Further analysis of simple main effects revealed that there was a significant difference between the modes for the hard-to-find images, $F(2, 23) = 3.76$, $p < 0.039$, as opposed to the easy-to-find images, $F(2, 23) = 2.02$, $p < 0.16$.

The participants using the eye tracking interface found the target in fewer steps than the automated random selection strategy and the analysis of simple effects attributed the significant difference to the hard-to-find images. This meant that the probability of finding the hard-to-find images was significantly increased due to human cognitive abilities as opposed to the indiscriminate selection by random selection.

There were many occasions when the fixations returned twice or more to the images that were finally selected, as shown in figure 5. The figure also shows that search becomes more directed with less opportunity for revisits as the gaze time is reduced from 800 ms to 400 ms.

4. Experiment 2

4.1 *Objective*

The first experiment demonstrated the feasibility of driving an image retrieval engine with an eye gaze interface. Analysis of the eye movement data during image search in the last section revealed that users frequently revisit images that are subsequently selected and do this quite happily at rapid speeds (400 ms cumulative fixation threshold). The objective of this experiment was to investigate the likelihood of reaching the target image using alternative criteria for improved image selection. The effects of lower fixation thresholds and revisits were investigated.

4.2 *Experiment design*

Twenty-four unpaid participants (18 males and six females) took part in this experiment. The mean age was 28.8 years with a median of 27 and a mode of 26. The same database of images and its pre-computed similarity links were used in this experiment. Eight easy-to-find target images were selected. The display automatically changed based on gaze behaviour. In this experiment, four treatments were used for determining best image selection:

- a cumulative fixation threshold of 400 ms as before;
- a shorter cumulative fixation threshold of 300 ms;
- selection by revisit; and
- selection by revisit or cumulative fixation threshold of 400 ms.

As in prior experiments the cumulative fixation threshold is determined by the accumulation of all fixations greater than 80 ms on a specific image position exceeding a 300 ms or 400 ms threshold. A revisit is determined by the re-fixation of an item that has been previously fixated. In this case the first image to be visited twice is selected as the selected image. The fourth treatment is determined by a revisit or the cumulative fixation threshold of 400 ms, whichever occurs first (i.e. the selected image is either determined by the first image revisited or the first image to exceed the cumulative threshold of 400 ms).

Results in the previous section showed that the inclusion of random images in successive displays did not affect performance. In this experiment the display used either:

(a) 15 images with the highest similarity values to the selected image, or
(b) 15 images with the highest similarity values to the 15th ranked similar image to the selected image.

Theoretically, condition (b) should allow users to move more freely between clusters, but at the risk of moving away from the target.

Participants performed eight runs, using easy-to-find image types. There was one practice run to enable better understanding of the task and to equalize skill levels before the experiment. Participants understood that there would be a continuous change of display until they found the target, but did not know what determined the display change. Eight treatment combinations of the four fixation thresholds (400 ms, 300 ms, Revisit and Revisit/400 ms) and two ranking levels ((a) and (b)) were applied. Any sequence effect was minimized by randomly allocating each participant to 24 different sequences of the four fixation thresholds. There was a 1 min rest between runs. As before, the maximum number of steps to target was limited to 26 screen changes.

4.3 *Results and analysis*

Three dependent variables, the number of steps to target, the time to target (F_1), and the number of fixations (F_2) of 80 ms and above, were monitored and recorded during the experiment. Twenty-four dependent variables (eight each) were recorded for each participant. The average figures are presented in table 3.

One hundred and ninety-two ($=8 \times 24$) figures were entered for each dependent variable into a repeated measures ANOVA with two factors (fixation threshold and ranking). The main effect of the fixation threshold was not significant, $F(3, 69) = 0.44$, $p = 0.724$, with similar steps to target as shown in table 3. Paired comparisons of all fixation thresholds also showed no significant difference in steps to target. The analysis of the time and fixations per display revealed that there were significant differences in the time to target and number of fixations per display for all paired comparisons. More importantly, Revisit/400 ms took significantly

Table 3. Analysis of human eye behaviour on the interface (rounded-off mean figures).

Fixation threshold	A	B	C	D	E	F
300 ms	50.0%	17	17.9	1.08	53	3
400 ms	56.3%	18	28.1	1.63	86	5
Revisit	45.8%	16	37.7	2.35	99	6
Revisit/400 ms	52.1%	17	24.0	1.47	72	4

A = target not found (frequency); B = steps to target; C = time to target (seconds); D = average time per display; E = fixation numbers; F = average fixation numbers per display.

less time ($p = 0.023$) and fewer fixations ($p = 0.042$) than the 400 ms threshold for making decisions in each display.

Combining revisits with a fixation threshold (i.e. Revisit/400 ms) reduced the time spent on each display sequence without affecting the search efficiency (i.e. steps to target) compared with the 400 ms threshold. Remarkably, the results also reveal that users are able to locate target images at the 300 ms fixation threshold level with fewer average steps to target than the 400 ms threshold (table 3). Although there was no significant difference between the steps to target for the 300 ms and 400 ms ($p < 0.55$), there was a significant difference for the time to target ($p < 0.0001$) and fixation numbers ($p < 0.0001$) in each display.

The same treatment combinations experienced by all participants were applied to the random selection tool to obtain 192 dependent variables (steps to target). By combining the variables, 384 figures were entered into a mixed design multivariate ANOVA with four observations per cell and two factors (selection mode and ranking). The average figures are presented in table 4.

In summary, the results of the ANOVA revealed a main effect of the selection mode, $F(4, 43) = 5.434$, $p = 0.001$, with the eye (17) taking significantly fewer steps than random selection (21). Conducting univariate tests on all four fixation threshold treatments with the corresponding random values generated by the random selection strategy revealed significant differences between the eye gaze and the random values for each fixation threshold treatment as follows:

- 300 ms, $F(1, 46) = 5.218$, $p = 0.027$;
- 400 ms, $F(1, 46) = 4.152$, $p = 0.047$;
- Revisit, $F(1, 46) = 8.107$, $p = 0.007$;
- Revisit/400 ms, $F(1, 46) = 5.730$, $p = 0.021$.

4.4 *Extended experiment*

An outstanding question is whether there is a limit to the speed of operation of this interface because users appear to obtain good performance at both 300 ms and 400 ms fixation thresholds. Therefore, an extended experiment was devised to investigate three cumulative fixation threshold levels of 300 ms, 200 ms and 100 ms. Three of the easy-to-find target images from the previous experiment were selected for this experiment. The choice of targets was based on the target images with the least average steps to target.

Six unpaid participants (four males and two females) took part in this experiment. The average age was 36.2 years. Each participant performed three runs using easy-to-find image types. Three treatment combinations of the three fixation thresholds (300 ms, 200 ms and 100 ms) were applied for each participant. Any sequence effect was minimized by randomly allocating each participant to six different sequences of the three fixation thresholds.

4.4.1 Results Three dependent variables, the number of steps to target, the time to target (F_1), and the number of fixations (F_2) of 80 ms and above, were monitored and recorded during

Table 4. Comparison of eye and random selection (rounded-off mean figures).

Selection mode	Target not found (frequency)	Steps to target
Eye gaze	51.0%	17
Random selection	69.8%	21

Table 5. Analysis of human eye behaviour on the interface (rounded-off mean figures).

Fixation threshold	A	B	C	D	E
100 ms	20	7.97	0.394	20	1
200 ms	12	6.96	0.634	18	2
300 ms	4	5.20	1.139	17	3

A = steps to target; B = time to target (seconds); C = average time per display; D = fixation numbers; E = average fixation numbers per display.

the experiment. Nine dependent variables (three each) were recorded for each participant. The average figures are presented in table 5.

Eighteen ($=3 \times 6$) figures were entered for each dependent variable into a single factor ANOVA with three levels (300 ms, 200 ms and 100 ms). The results of the ANOVA performed on the steps to target revealed a significant main effect of the fixation thresholds, $F(2, 10) = 13.098$, $p = 0.018$. A paired comparison of 100 ms and 300 ms attributed the significant difference to a simple main effect between these two fixation thresholds ($p = 0.003$). There was no significant difference between the 100 ms and 200 ms paired thresholds ($p = 0.133$) and 200 ms and 300 ms paired thresholds ($p = 0.227$), respectively.

The same treatment combinations experienced by all participants were applied to the random selection tool to obtain 18 dependent variables (steps to target). By combining the variables, 36 figures were entered into a multivariate ANOVA with three observations per cell and one factor (selection mode). In summary, the results of the ANOVA revealed a main effect of the selection mode, $F(3, 8) = 6.348$, $p = 0.016$. The eye (12) took significantly fewer steps to the target than the random selection (21). Univariate tests on all three fixation threshold levels with the corresponding values generated by the random selection strategy revealed significant differences between the eye gaze and random values for the 300 ms and 200 ms conditions, i.e. $F(1, 10) = 10.390$, $p = 0.009$ and $F(1, 10) = 9.484$, $p = 0.012$, respectively. At the cumulative threshold level of 100 ms, participant's gaze behaviour effectively became randomized, $F(1, 10) = 0.056$, $p = 0.817$. The performance at the 300 ms threshold and certainly the 200 ms threshold indicated that pre-attentive vision was being employed by participants to obtain these response times [1].

5. Conclusions

Experiments have shown that an eye tracking image retrieval interface together with pre-computed similarity measures yield a significantly better performance than random selection using the same similarity information. A significant effect on performance was also observed with hard-to-find images. This was not seen with easy-to-find images where, with the current database size, a random search might be expected to perform well.

Additional experiments have revealed that re-fixation or revisits on an image may be an indication of interest in an image. Furthermore, participants were able to find target images with a 200 ms fixation threshold, indicating that rapid pre-attentive vision was being used in the experiments. Improving performance and decreasing costs will mean that, before long, eye trackers will reach the mass market where they will replace existing interfaces in which vision and search currently play a major part. In addition, eye trackers are certain to be an essential tool in most aspects of research into human vision.

Acknowledgements

The authors acknowledge the support of BT Research and Venturing, SIRA and the Engineering and Physical Sciences Research Council for this work. The work was conducted within the framework of the EC-funded Network of Excellence (MUSCLE) [20].

References

[1] Venters, C.C., Eakins, J.P. and Hartley, R.J., 1997, The user interface and content based image retrieval systems. Paper presented at the 19th BCS-IRSG Research Colloquium, Aberdeen, April.
[2] Duchowski, A.T., 2002, A breadth-first survey of eye tracking applications. *Behaviour Research Methods, Instruments, & Computers (BRMIC)*, **34**, 455–470.
[3] Hansen, J.P., Anderson, A.W. and Roed, P., 1995, Eye gaze control of multimedia systems. In: Y. Anzai, K. Ogawa and H. Mori (Eds) *Symbiosis of Human and Artifact*, Vol 20A (Amsterdam: Elsevier Science), pp. 37–42.
[4] Numajiri, T., Nakamura, A. and Kuno, Y., 2002, Speed browser controlled by eye movements. Paper presented at the IEEE International Conference on Multimedia and Expo, Lausanne, 26–29 August.
[5] Corno, F., Farinetti, L. and Signorile, I., 2002, A cost effective solution for eye-gaze assistive technology. Paper presented at the IEEE International Conference on Multimedia and Expo, Lausanne, 26–29 August.
[6] Ward, D.J. and MacKay, D.J.C., 2002, Fast hands-free writing by gaze direction. *Nature*, **418**, 838.
[7] McCarthy, J., Sasse, M.A. and Riegelsberger, J., 2003, Could I have the menu please? An eye tracking study of design conventions. Paper presented at HCI2003, Bath, UK, 8–12 September.
[8] Henderson, J.M. and Hollingworth, A., 1999, High-level scene perception. *Annual Reviews Psychology*, **50**, 243–71.
[9] Puolamki, K., Salojrvi, J., Savia, E., Simola, J. and Kaski, S., 2005, Combining eye movements and collaborative filtering for proactive information retrieval. Paper presented at the 28th ACM Conference on Research and Development in Information Retrieval (SIGIR).
[10] Itti, L., 2004, Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, **13**, 1304–1318.
[11] Cox, I.J., Miller, M.L., Minka, T.P., Papathomas, T.V. and Yianilos, P.N., 2000, The Bayesian image retrieval system, PicHunter: theory, implementation, and Psychophysical experiments. *IEEE Transactions on Image Processing*, **9**.
[12] Yarbus, A., 1967, *Eye Movements and Vision* (New York: Plenum Press).
[13] Mackworth, N. and Morandi, A., 1967, The gaze selects informative details within pictures. *Perception and Psychophysics*, **2**, 547–552.
[14] Oyekoya, O.K. and Stentiford, F.W.M., 2004, Exploring human eye behaviour using a model of visual attention. In: *Proceedings of the 17th International Conference on Pattern Recognition*, ICPR 2004, Vol. 4, pp. 945–948.
[15] Oyekoya, O.K. and Stentiford, F.W.M., 2005, A performance comparison of eye tracking and mouse interfaces in a target image identification task. In: *Proceedings of the 2nd European Workshop on the Integration of Knowledge, Semantics & Digital Media Technology*, EWIMT 2005, pp. 139–144.
[16] Oyekoya, O. and Stentiford, F., 2006, An eye tracking interface for image search. In: *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications*, ETRA 2006, pp. 40–40.
[17] LC Technologies Inc., http://www.eyegaze.com/.
[18] Stentiford, F.W.M., 2007, Attention based similarity. *Pattern Recognition*, **7**, 771–783.
[19] Fitts, P.M., Jones, R.E. and Milton, J.L., 1950, Eye movement of aircraft pilots during instrument-landing approaches. *Aeronautical Engineering Review*, **9**, 24–29.
[20] Multimedia Understanding Through Semantics, Computation and Learning, Network of Excellence, EC 6th Framework Programme, FP6-507752, http://www.muscle-noe.org/.
[21] Treisman, A.M. and Gelade, G., 1980, A feature-integration theory of attention. *Cognitive Psychology*, **12**, 97–136.