Journal of Eye Movement Research
2(2):3. 1-15

Fixation sequences in imagery and in recognition during the processing of pictures of real-world scenes

Katherine Humphrey
University of Nottingham

Geoffrey Underwood
University of Nottingham

Thirty photographs of real-world scenes were presented for encoding, and half the participants then performed a recognition test, deciding whether each of 60 images were old (from the original set) or new. The other participants performed an imagery task immediately after encoding each of the 30 images. After completing this task, the recognition group then performed the imagery task in response to prompts that were unique verbal descriptors, and the imagery group performed the recognition task. All participants returned 2 days later, and repeated the imagery test. Eye movements were recorded during all phases.

Differences in average fixation duration, average number of fixations and average saccadic amplitude were found between task groups and between experimental phases. Scan patterns were compared with a string-editing algorithm. Close similarities were observed between experimental phases that involved more similar tasks (e.g., initial *encoding vs. recognition*, and *immediate imagery vs. delayed imagery*). Scan patterns were equally similar when the task was presented immediately or after 2 days. We propose that the more similar the encoding and retrieval processes are, the more similar eye movements will be at each of these experimental stages.

Keywords: Eye movements, Imagery, Scan pattern, Visual Buffer

Introduction

The order and pattern of fixations and saccades made by the viewer when looking at a scene has been described as a 'scanpath' by Noton and Stark (1971a,b) in a theory that predicts that the fixations made when first looking at a picture are very similar to those they make when recognising that picture at a later time. Scanpath Theory makes unsupported assumptions about the neural mechanisms that result in re-instated sequences, arguing that the sequence becomes part of the memory of the picture and that the oculomotor pattern becomes part of an integrated representation in memory, and as a result the theory continues to attract criticism (e.g., Henderson, 2003). Repeated viewing of an image does result in a sequence of fixations that is similar to the sequence made during the first inspection, but it is questionable whether the mechanism of repetition involves an integrated perceptual-motor representation. Accordingly, we will avoid the term 'scanpath' here, and opt for the more cautious 'scan patterns' in describing sequences of fixations.

A number of studies have found that when participants view a picture for the second time, the scan patterns they produce are very similar to scan patterns produced on first exposure to the picture. For example, in Foulsham and Underwood's (2008) recognition memory study participants first inspected a set of 45 pictures. They were then shown another set of 90 pictures and were asked to decide whether they had seen each picture before. It was found that scan patterns were most similar when compared between two viewings of the same picture (encoding *vs.* old). The similarity was significantly greater than control comparisons (encoding *vs.* new and old *vs.* new). However, an argument against Scanpath Theory is that people may not reproduce the same scan patterns over time due to the sequence of eye movements being stored internally or being related to an internal visual image, but they do so by chance because of the bottom-up influences

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

of the visual stimulus. When we view a picture (at least in a free-viewing or in a memory task), our eye fixations are attracted by the visual saliency of the image, with more attention being given to conspicuous regions than elsewhere. When we are shown that same picture again at a later time, perhaps we simply look at the same parts of the picture again, as those parts still hold the same low-level properties as when it was first inspected. By this argument, the re-instatement of a sequence of fixations on separate occasions may be a product of the visual characteristics of the image rather than having any involvement with our memories of the image or of our scan pattern on first viewing.

Similar to Scanpath Theory is the Perceptual Activation (PA) Theory (e.g. Thomas, 1999). According to PA, we are able to examine, explore and interpret a scene because of the continual updating and refining of procedures (or "schemata" [Neisser, 1976]) that specify how to direct our attention. However, there are no stored descriptions or pictures. This theory is similar to, but avoids many of the criticisms of Scanpath Theory as no *thing* in the brain *is* the percept or image.

Saliency (an item's quality of being visually distinctive relative to its neighboring items) has been shown to affect the order and pattern of fixation. Koch and Ullman (1985) and Itti and Koch (2000) proposed that attention is drawn to the most salient region in an image first, followed by the second most salient region then the third most salient region, and so on. Attention, and eye fixations, are attracted to the region identified as being of greatest brightness, colour contrast and orientation change, and once we have fixated that region a process of inhibition of return prevents attention from being locked onto any one region, and allows us to saccade to the next most salient region.

The potency of saliency has been demonstrated in a number of studies. For example, Sheth and Shimojo (2001) briefly displayed a target and then asked participants to point to its previous location. Participants estimated targets to be closer to the centre of gaze, and closer to visually salient markers in the visual display than they actually were. The locations of objects presented earlier were *remembered* falsely as being closer to salient reference frames than they really were. Salient regions attract fixations when viewers are not given an explicit purpose in looking at a picture. Parkhurst, Law, and Niebur (2002), showed viewers a range of images and recorded eye movements. Saliency strongly predicted fixation probability during the first two or three fixations, and the model performed above chance throughout each trial. In contrast to this, Tatler, Baddeley and Gilchrist (2005) found no change in the involvement of image features over time and Tatler (2007) argues that even the *correlation* between features and fixations is minimal.

Further support for a saliency map model of scene inspection comes from Underwood, Foulsham, van Loon, Humphreys, and Bloyce, (2006) and from Underwood and Foulsham (2006), who found that when viewers inspected the scene in preparation for a memory task, objects higher in saliency were potent in attracting early fixations. These studies of the effects of saliency could suggest that scan patterns are similar at encoding and recognition not because of an internally stored sequence of fixations, but because the same bottom-up features are present at both encoding and recognition, and therefore participants just look at the same conspicuous parts of the scene.

There is evidence that bottom-up saliency can be overridden by top-down knowledge (Humphrey and Underwood, 2008) and by task variations that emphasise the search for specific characteristics (Underwood et al., 2006; Underwood & Foulsham, 2006; Underwood, Templeman, Lamming and Foulsham, 2008). Henderson, Brockmole, Castelhano, and Mack (2007) found that during an active search task, neither region-to-region saccades nor saccade sequences were predicted any better by visual saliency than by a random model. There were differences in intensity, contrast, and edge density at fixated scene regions compared to regions that were not fixated, but these fixated regions also differed in rated semantic informativeness. Similarly, Einhäuser, Rutishauser, and Koch, (2008) found that during free-viewing, observers' eye-positions were immediately biased toward the high-saliency side of a picture. However, this sensory driven bias disappeared entirely when observers searched for a target embedded with equal probability to either side of the stimulus. When the target always occurred in the low-contrast side, observers' eye-positions were immediately biased towards this low-saliency side, i.e., the sensory-driven bias reversed.

Even when saliency is overridden by the task demands, it could still be argued that scan patterns are reproduced because the same semantically interesting parts of the scene are present at encoding and recognition. Re-

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

peated scan patterns may be generated by viewers remembering how they inspected a picture when they first looked at it, but it could be that the features of the image - either bottom-up visual features or top-down meaningful features – are what drive the sequence of fixations.

One way to get around these problems is to use an imagery task, so that if scan patterns are reproduced, it cannot be due to external bottom-up influences, as no visual stimulus is present. Brandt and Stark (1997) found substantial similarities between sequences of fixations made whilst viewing a simple checker-board diagram and those made when imagining it later. Since there is no actual diagram or picture to be seen during the imagery period, it is likely that an internalised cognitive perceptual model is in control of these scan patterns. Holsanova, Hedberg and Nilsson (1998) used natural, real life scenes and found results similar to those reported by Brandt and Stark.

In a modified version of the imagery experiment, Laeng and Teodorescu (2002) manipulated when participants could move their eyes. Participants that were told to keep their eyes centrally fixated during the initial scene perception did the same, spontaneously, during imagery. Participants that were allowed to move their eyes during initial perception but were told to keep their eyes centrally fixated during imagery exhibited decreased ability to recall the pattern. Laeng and Teodorescu argued that this was because the oculomotor links established during perception could not be used in the process of building up a mental image, and this limitation impaired recall. Eye movements at first viewing help to encode the picture and reproducing those eye movements at a later stage may help recall the picture. However, it could be argued that when pictures are better recalled, the eye movement patterns during imagery, as a *result*, better match the eye movement patterns during scene viewing. A decrease in recall performance when participants are instructed to keep fixation at imagery could therefore be due to additional cognitive load exhibited by the (additional) task to refrain from naturally moving one's eyes.

One aim of the current paper is to determine whether scan patterns are reproduced during imagery. This could avoid the criticisms that the reproduction of scan patterns may be due to external bottom-up influences, as this cannot be true if no visual stimulus is present.

It would be interesting to know whether this relationship between imagery and perception persists over time. Ishai and Sagi (1995) have shown, for example, that imagery induced facilitation in a target-detection task decays and is only effective in the first 5 min after the participants saw the stimuli. In Laeng and Teodorescu's (2002) study, the participants performed the imagery task 40 seconds after they studied the stimuli and it was suggested by Mast and Kosslyn (2002) that the sensorimotor trace may be stored only in short-term memory. One aim of the current experiment is to determine whether scan patterns at imagery are stable over extended periods of time.

One model that could help explain eye movements during imagery is Kosslyn's (1994) 'visual buffer', which is used to construct an internal image. The visual buffer is located in the working memory, which is topographically organized and has the possibility to represent spatiality. An 'attention window' can be moved to certain parts of the visual buffer, which could be connected to eye movements during imagery. Mental images are generated in the visual buffer, and representations of those images are stored in long term memory. When a scene stored in long term memory is visualized, it is generated (or rather created or re-created) in the working memory and in the visual buffer.

A large amount of criticism against the visual buffer comes from propositional accounts (e.g., Pylyshyn, 2002, 2003), which claim that there are no such things as internal images. Pylyshyn argues that imagined objects and spatial locations are bound to visual features in the external world; these bindings are called 'visual indexes' (Pylyshyn, 2000, 2001, 2002). This theory assumes no pictorial properties whatsoever of the 'projected image', only the binding of imagined objects to real, perceived ones. However, Johansson, Holsanova and Holmqvist (2006) carried out an imagery study in the dark (i.e., without any possible visual features) and still yielded eye movements that reflected objects from both the description and the picture. Therefore, Johansson et al. argued that visual indexes that only assume the binding of propositional objects to real ones cannot explain eye movements during mental imagery.

One aim of this experiment is to investigate which account best explains eye movements during imagery, and also whether eye movements at retrieval are affected by different methods of encoding and of retrieval. If Pyly-

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

shyn's propositional model holds true, then eye movements should not be affected by such manipulations, as they would not change tacit knowledge (the knowledge of what seeing a specific object would be like). This study also aims to find out if, assumed that a scan pattern is reproduced; temporal information is reproduced as well as spatial information. To do this, average fixation duration, average saccadic amplitude, and the number of fixations are calculated at each encoding and retrieval condition. Two procedures were used in the experiment, one in which viewers were required to visualize the picture most recently inspected, and one in which the imagery task was conducted after the presentation of all of the pictures in the experiment. In both procedures there was an imagery task and a recognition memory task – the order was reversed between procedures. After a two day interval the imagery task was repeated.

Method

Participants

Thirty participants took part in the experiments, all of whom were students at Nottingham University. The age range was 18-51 and the mean age was 25.5. The sample comprised 21 females and 9 males. All participants had normal or corrected-to-normal vision. Inclusion in the study was contingent on reliable eye tracking calibration and the participants being naïve to eye movements being recorded.

Materials and apparatus

Eye position was recorded using an SMI iVIEW X Hi-Speed eye tracker, which uses an ergonomic chinrest and provides very precise data within a gaze position accuracy of 0.2 degrees. The system parses samples into fixations and saccades based on velocity across samples, with a spatial resolution of 0.01°, a processing latency of less than 0.5 milliseconds and a sampling rate of 240 Hz. A set of 60 high-resolution digital photographs were prepared as stimuli, sourced from a commercially available CD-ROM collection and taken using a 5MP digital camera. Each picture was distinctly individual, in that given a short sentence describing a picture; it could not be mistaken for any of the others. Examples of these stimuli are shown in Figure 1.

A pilot study was conducted to make sure the stimuli were distinctly individual and could not be confused. Ten participants were given a sheet of 60 pictures and a sheet of 60 descriptive labels, both randomly ordered, and were asked to match the pictures to the labels. All of the participants correctly matched 100% of the stimuli.

Half of each category were designated "old" and shown in both encoding and test phases, while the other half were labelled "new" and were shown only as fillers at test. New and old pictures were similar in complexity, semantic and emotional content. Pictures were presented on a colour computer monitor at a resolution of 1600 by 1200 pixels. The monitor measured 43.5cm by 32.5cm, and a fixed viewing distance of 98cm gave an image that subtended 25.03 by 18.83 degrees of visual angle.

Figure 1: Examples of two of the distinctively individual pictorial stimuli used in the experiments: 'the penguins', and 'the buttons'.

Design

The experiment used a between groups design, with 2 groups of participants (15 participants in each group). The independent variable was therefore which group the participant belonged to (The Imagery First group or The Recognition First group). The dependent variable measures were: accuracy in deciding whether a picture was old or new, average fixation durations, average saccadic amplitude, average number of fixations, and the similarities of scan patterns compared at encoding and imagery, encoding and recognition, encoding and delayed imagery, imagery and recognition, imagery and delayed recognition, and recognition and delayed imagery.

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

Procedure

Participants were told that their pupil size was being measured in relation to mental workload. They were informed that although their eye movements were not being recorded, it was important to keep their eyes open so pupil size could be reliably measured.

Task1: Imagery Prior to Recognition. Following a 9-point calibration procedure, participants were shown written instructions on the experimental procedure and given a short practice. The first stage involved seeing a picture for 3000 milliseconds then a brightly coloured mask for 1000 milliseconds and then the screen turned blank. The participant then had 5000 milliseconds to visualize the last photograph they had seen. After this time, a fixation cross appeared for 1000 milliseconds to ensure that fixation at picture onset was in the centre of the screen. This experimental procedure is illustrated in figure 2.

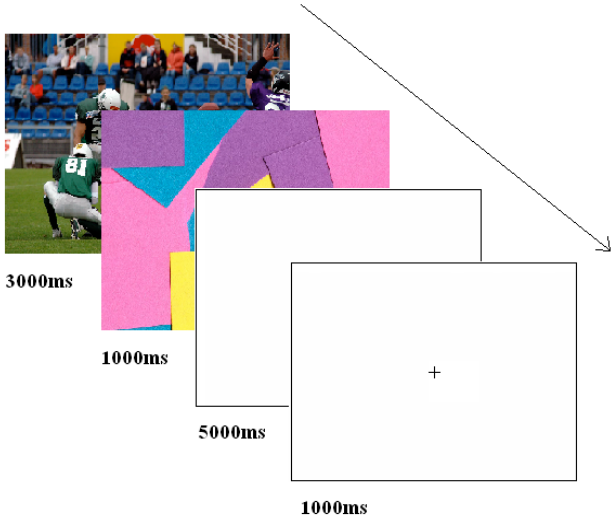

Figure 2: A diagrammatic representation of the imagery prior to recognition procedure.

After participants had seen and visualized 30 stimuli, presented in a random order, they took a short break and were then asked to perform a recognition memory test. Participants saw a second set of pictures and had to decide whether each picture was new (never seen before) or old (from the previous set of pictures). They were instructed to press "N" on the keyboard if the picture was new, and "O" on the keyboard of the picture was old. Sixty stimuli were presented in a random order, 30 of which were old and 30 new. In order to facilitate an ideal comparison of scan patterns between encoding and recognition, each picture was shown for 3000 milliseconds and participants could only make a response after this time. This was to encourage scanning of the whole picture. This procedure is illustrated in figure 3.

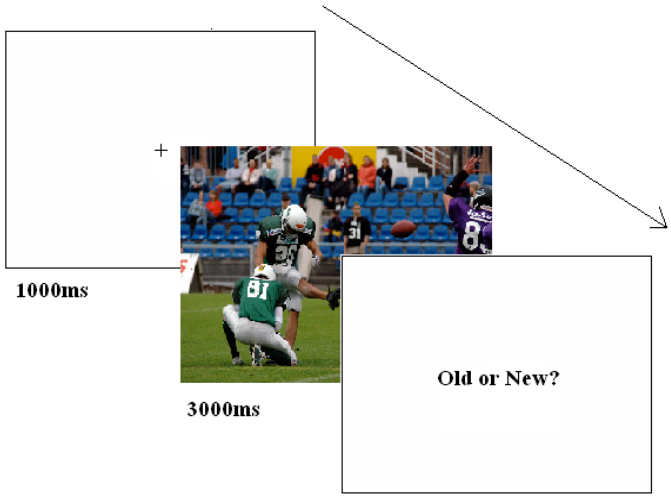

Figure 3: A diagrammatic representation of the recognition memory test in the 'imagery prior to recognition' procedure.

Participants returned approximately 48 hours later to perform another imagery task. This time they saw 30 white screens with a short sentence describing one of the pictures seen 48 hours earlier. All of the pictures described in this task had previously appeared in the first imagery task, and were presented here in a new random sequence. Participants were asked to visualize the picture described and try to remember everything they could about it. Each description appeared for 3000 milliseconds and then the screen went blank for 5000 milliseconds, during which they visualized the stimulus. This procedure is illustrated in figure 4.

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

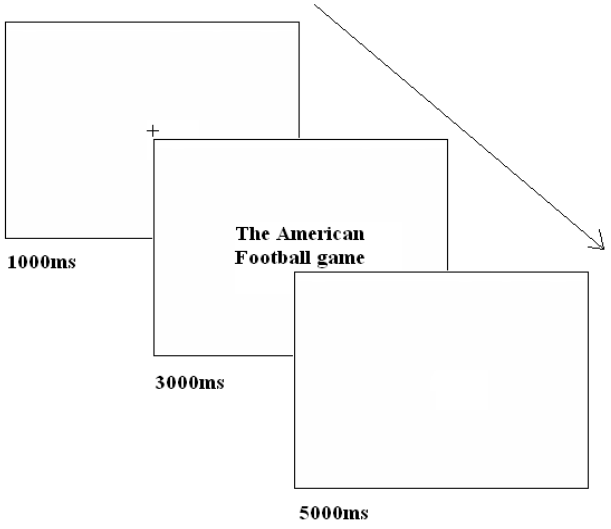

Figure 4: A diagrammatic representation of the delayed imagery task in both the 'imagery prior to recognition' and the 'recognition prior to imagery' procedures.

Task 2: Recognition Prior to Imagery. The difference between The Imagery First group and The Recognition First group was in the order of the imagery and recognition tasks. As before, the first stage here involved viewing a set of 30 stimuli, presented in a random order, in preparation for a memory test, but no imagery took place at this stage. Each picture was preceded by a fixation cross for 1000 milliseconds, which ensured that fixation at picture onset was in the centre of the screen. Each picture was presented for 3000 milliseconds, during which time participants moved their eyes freely around the screen.

After all 30 pictures had been presented, participants saw a second set of pictures and had to decide whether each picture was new (not seen before in the experiment) or old (from the previous set of pictures). They were instructed to press "N" on the keyboard if the picture was new, and "O" on the keyboard of the picture was old. Sixty stimuli were presented in a random order, 30 of which were old and 30 new. In order to facilitate an ideal comparison of scan patterns between encoding and recognition, each picture was shown for 3000 milliseconds and participants could only make a response after this time. This was to encourage scanning of the whole picture. See figure 3 for a diagrammatic illustration of this recognition procedure.

After all 60 pictures in the recognition test had been shown, the participants took a break before performing an imagery task. This time they saw 30 white screens with a short sentence describing one of the pictures they had just seen. All the pictures in this imagery task were classified as 'old' but the participants were not informed of this. The pictures appeared in a random order. Participants were asked to visualize the picture described and try to remember everything they could about it. Each stimulus appeared for 3000 milliseconds and then the screen went blank for 5000 milliseconds, in which they visualized the stimulus. See figure 4 for a diagrammatic illustration of this procedure.

Participants returned two days later to perform the last imagery task again (see figure 4). The procedure was identical and all of the descriptions of pictures in this task had previously appeared in the first imagery task, and were presented here in a new random order. Participants were asked to visualize the picture described and try to remember everything they could about it. Each description appeared for 3000 milliseconds and then the screen went blank for 5000 milliseconds, in which time they visualized the stimulus.

Results

In all cases, trials were excluded where the fixation at picture onset was not within the central region (the central square around the fixation cross when the picture was split into a 5x5 grid at analysis), or when calibration was temporarily interrupted (e.g. if the participant sneezed, therefore removing their head from the eye tracker).

There were 2 main types of data, recognition memory data (accuracy), and eye tracking measures – average fixation durations, average saccadic amplitude, average number of fixations, and string analyses.

Although participants in both Tasks performed both the imagery and recognition tests but in different orders, for the sake of clarity Task 1 will be referred to as the 'Imagery First group' and Task 2 will be referred to as the 'Recognition First group'.

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

At the end of both Tasks, participants filled out a short questionnaire consisting of 9 filler questions (e.g. age, degree course, level of tiredness etc) and one target question asking them about the aim of the experiment. One participant in the Imagery First group guessed the aim of the study and their data was discarded.

Recognition Memory

Accuracy. Accuracy was measured by the number of pictures participants correctly identified as 'old' (if they were from the previous set) or 'new' (if they had never been seen before). As shown in Figure 5, both groups performed at a very high accuracy rate (98.10% in the Imagery First group and 97.11% in the Recognition first group).

Data from one participant in The Imagery First group had to be removed because they pressed the wrong button all the way through the recognition test. A between-groups t-test on the remaining 28 participants showed no reliable difference between the groups: t(26)=0.97, p=0.623.

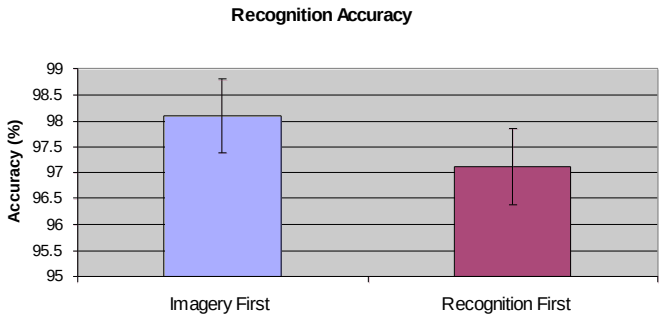

Figure 5: The mean recognition accuracies in the two Tasks. In The Imagery First group, participants attempted to visualize each picture immediately after viewing it, and in The Recognition First group they performed the imagery task after a recognition test.

Eye-tracking measures

Average Fixation Duration. Overall, participants in The Imagery First group exhibited shorter fixations than participants in The Recognition First group. These means are shown in Figure 6. In both Tasks, participants made shorter fixations at encoding than at imagery or delayed imagery. Participants also made shorter fixations at recognition (old and new pictures) than at imagery or delayed imagery.

A mixed-design ANOVA showed a reliable effect of group (Imagery First or Recognition First), $F(1,27) = 17.89$, $MSe = 128692$, $p<0.001$, and a reliable effect of test phase, $F(4,27) = 45.39$, $MSe = 128692$, $p<0.001$.

A post-hoc t-test indicated that fixation durations were shorter in The Imagery First group than in The Recognition First group ($t = 4.23$, $p<0.001$). Fixations were also shorter during encoding than during the first imagery phase ($t = 7.69$, $p<0.001$), and during the delayed imagery phase ($t = 7.74$, $p<0.001$). There were also differences between the imagery phases and the viewing of pictures during recognition: there were shorter fixations on old pictures ($t = 8.76$, $p<0.001$) and on new pictures during the recognition phase ($t = 9.39$, $p<0.001$), relative to the initial imagery phase. Similarly, there were shorter fixations on old pictures ($t = 8.81$, $p<0.001$) and on new pictures ($t = 9.45$, $p<0.001$) relative to fixation during the delayed imagery phase.

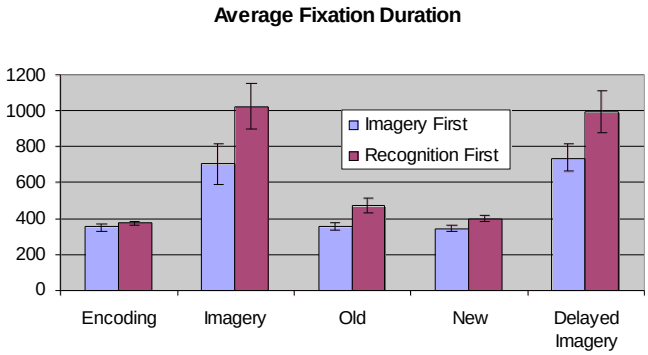

Figure 6: Differences in average fixation durations between the two Tasks and between phases of the course of the experiment.

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

Average Number of Fixations. The mean numbers of fixations made in each phase of the experiment and for each group of participants, are shown in Figure 7. Participants made more fixations at encoding then at imagery or delayed imagery. Participants also made fewer fixations at imagery and at delayed imagery than at recognition (old and new).

A mixed design ANOVA showed a reliable effect of test phase $F(4,27) = 20.10$, $MSe = 9328.043$, $p<0.001$.

Post-hoc t-tests showed that there were more fixations during encoding than during imagery ($t = 5.80$, $p<0.001$), and delayed imagery ($t = 5.90$, $p<0.001$). There were also more fixations on old pictures during the recognition phase than there were during the initial imagery phase ($t = 4.95$, $p<0.001$) or during delayed imagery ($t = 5.05$, $p<0.001$), and there were more fixations on new pictures ($t = 6.26$, $p<0.001$) and on old pictures ($t = 6.36$, $p<0.001$), than during the delayed imagery phase.

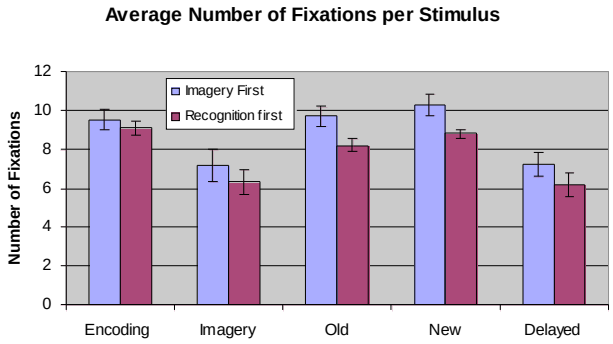

Figure 7: Differences in number of fixations between The Imagery First group and The Recognition First group and between phases of the course of the experiment.

Average Saccadic Amplitude. The average saccadic amplitudes in each phase of the experiment and for each group of participants are shown in Figure 8. Participants in the Recognition First group produced greater saccadic amplitudes than participants in the Imagery First group. Participants also produced greater saccadic amplitudes at imagery than at encoding and at delayed imagery than at encoding.

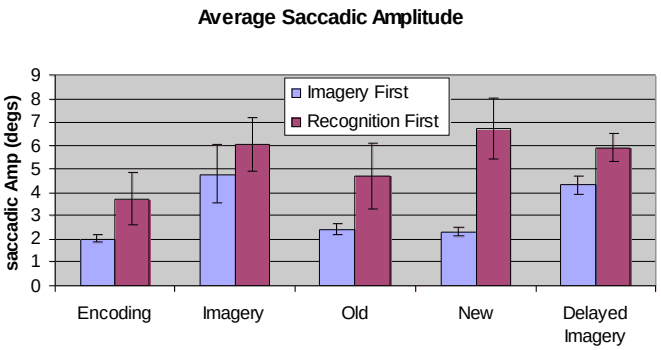

Figure 8: Differences in average saccadic amplitude between The Imagery First group and The Recognition First group and between phases of the course of the experiment.

A mixed design ANOVA showed a reliable effect of group (Imagery First or Recognition First) $F(1,27) = 13.987$, $MSe = 3795.602$, $p<0.001$. There was also a reliable effect of test phase $F(4,27)= 2.640$, $MSe = 3795.602$, $p<0.05$.

Post-hoc t-tests showed that there were reliable differences between encoding and imagery (t=2.73), and between encoding and delayed imagery (t=2.43).

Scan Patterns: String Editing

String editing was used to analyse the similarity between scan patterns produced on encoding and imagery, encoding and recognition, encoding and delayed imagery, imagery and recognition, imagery and delayed recognition, and recognition and delayed imagery. This string editing technique is described in detail by Brandt and Stark (1997); Choi, Mosley, & Stark, (1995); Hacisalihzade, Allen, and Stark, (1992), Privitera, Stark and Zangemeister (2007) and Foulsham and Underwood (2008) and involves turning a sequence of fixations into a string of characters by segregating the stimulus into labelled regions. The similarity between two strings is then computed by calculating the minimum number of editing steps required to turn one into the other. Three types of operations are permitted: insertions, deletions and substitutions. Similarity is given by one minus the number of edits required, standardised over the length of the string.

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

An algorithm for calculating the minimum editing cost is given in Brandt and Stark (1997) and this was implemented in the present study.

In the present study a 5 by 5 grid was overlaid onto the stimuli (see Figure 9). The resulting 25 regions were labelled with the characters A to Y from left to right. Fixations were then labelled automatically by the program, according to their spatial coordinates, resulting in a character string representing all the fixations made in this trial.

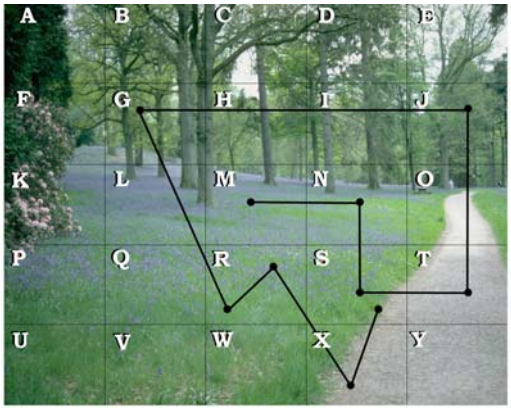

Figure 9: an example of the 5-by-5 grid on top of a photograph. An example fixation sequence has been drawn, with the initial fixation in the centre of the picture (grid box M).

For the fixation sequence shown in Figure 9, the string would be MNSTJGRRXS. The first fixation, which was always in the centre or region "M", was removed and adjacent fixations on the same regions were condensed into one (making the example NSTJGRXS). Repetitions were condensed because it is the global movements that are of interest here, rather than the small re-adjustments which combine to give one gaze on a region. Once the strings had been produced for all trials, they were compared using the editing algorithm and an average string similarity was produced across trials.

In our previous string editing analyses, strings were cropped to five letters to provide standardised and manageable data sets that were still long enough to display any emerging similarity (Foulsham and Underwood, 2008). However, the average number of fixations made by participants in the current experiment was eleven, so to test which string length was most appropriate, analyses conducted using 5-letter strings were compared to the same analyses using 11-letter strings. As t-tests showed no statistically reliable differences based on the average number of fixations included, strings were cropped to eleven letters for the following scan pattern analyses. In those trials where fewer than eleven fixations remained after condensing gazes, the comparison strings were trimmed to the same length.

The results were compared against a chance baseline. One way we considered doing this was to compare the experimental data against a random model. For example if more human gazes than randomly generated gazes lie in salient regions then this would suggest the visual system is selecting based on saliency. However, a uniformly distributed random model might lead to a difference purely due to systematic bias in eye movements towards the centre (see Tatler et al, 2005). Therefore, for each picture a participant viewed, the scanpath produced was compared to a scanpath that the participant produced on another a randomly selected picture. This was repeated for all 30 participants and an average similarity of 0.1159 was calculated.

Several experiments have shown that subjects rotate, change size, change shape, change colour, and reorganize and reinterpret mental images (e.g. Finke, 1989; Johansson, Holsanova, and Holmqvist, 2006). Although this could be a potential problem for the current paper, it will also be interesting to see whether scan patterns (and saccadic amplitudes) are highly similar at imagery and delayed imagery, suggesting that the reorganisation occurs mostly between encoding and imagery but then stays relatively stable over multiple imagery tasks.

The results of the comparisons are shown in Figure 10. In the Imagery First group eye movements were more similar when comparing imagery and delayed imagery than when comparing encoding and imagery or encoding and delayed imagery or Imagery and recognition. Example scan patterns from one participant in The Imagery First group (chosen at random) are also shown in figure 11 and compare encoding, imagery and recognition phases.

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

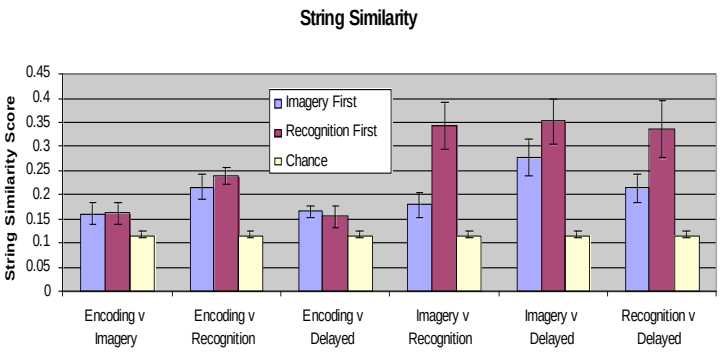

Figure 10: Differences in string similarities between The Imagery First group and The Recognition First group and between string comparison types.

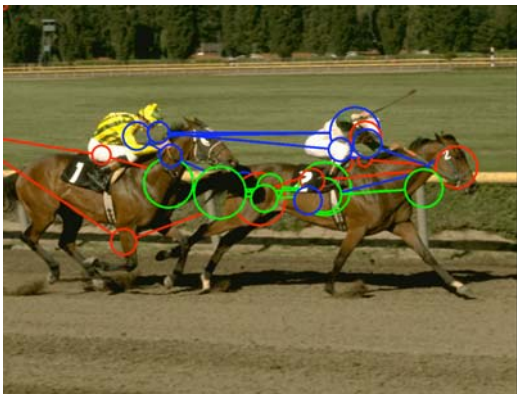

Figure 11: Example scan patterns from one participant in The Imagery First group, chosen at random. The blue scan pattern represents fixations and saccades at encoding; the red at imagery; and the blue at recognition. Encoding and recognition are slightly more similar than encoding and imagery or recognition and imagery. Note the increased saccadic amplitudes at imagery.

In The Recognition First group, eye movements were less similar when comparing encoding and imagery then when comparing encoding and recognition, imagery and recognition or imagery and delayed imagery. Eye movements were more similar when comparing encoding and recognition than when comparing encoding and delayed imagery. Eye movements were less similar when comparing encoding and recognition than when comparing imagery and recognition, imagery and delayed imagery or recognition and delayed imagery. Eye movements were less similar when comparing encoding and delayed than when comparing imagery and recognition or imagery and delayed or recognition and delayed.

A mixed design ANOVA showed a reliable effect of string comparison type: $F(5,27) = 11.23$, $MSe = 0.052$, $p<0.001$, and a reliable interaction between group (Imagery First and Recognition First) and string comparison type: $F(5,135) = 3.57$, $MSe = 0.010$, $p<0.01$. In the following tests we used the string similarity scores that are shown in Figure 10, and compared them against each other. To identify the source of the interaction, this was done for each of the Tasks. Because there are a large number of potential comparisons (30), only comparisons within phase-types will be considered here.

Post-hoc t-tests showed that for The Imagery First group, there were reliable differences between string similarities *encoding v imagery* and *imagery v delayed imagery* ($t = 3.04$, $p<0.01$), between *encoding v delayed imagery* and *imagery v delayed imagery* ($t = 2.96$, $p<0.01$), and between *imagery v recognition* and *imagery v delayed imagery* ($t = 2.59$, $p<0.05$). In each of these three comparisons, the *imagery v delayed imagery* similarity was the greater of the two.

In The Recognition First group, post-hoc t-tests showed that there were reliable differences between *encoding vs. imagery* and *encoding vs. recognition* ($t = 2.09$, $p<0.05$), with more similarity between scan patterns involving recognition than imagery. There were also differences between *encoding vs. imagery* and *recognition vs. imagery* ($t = 4.94$, $p<0.001$), and between *encoding vs. imagery* and *imagery vs. delayed imagery* ($t = 5.19$, $p<0.001$). In each of these comparisons the similarity of *encoding vs. imagery* had the smaller magnitude. As with The Imagery First group, the similarity score for *encoding vs. delayed imagery* was greater than that for *encoding vs. recognition* ($t = 2.28$, $p<0.05$).

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

Discussion

The main aims of this study were to determine whether scan patterns are reproduced when no visual stimulus is present and thus arguing against fixation selection being based on low level factors; to determine whether scan patterns at imagery are stable over time; to determine which account (Visual Buffer/propositional theory) best explains eye movements a imagery; and to determine whether eye movements at retrieval are affected by methods of encoding and of retrieval.

Analyses of recognition memory showed that participants in both Tasks were very good at identifying pictures as old or new. The accuracy was so high because each picture had to be distinctly individual in order for the imagery and delayed imagery tasks to work. This made it easy to decide which pictures had been seen before and which had not.

Average fixation durations were measured and analyses found a main effect of group in that participants in The Imagery First group made shorter fixations than participants in The Recognition First group. Average fixation duration at encoding was almost identical for The Imagery First group and The Recognition First group; suggesting that the groups were well matched and the differences between groups in other conditions were effects of the experimental design. This was also true of number of fixations. Interestingly, there was a difference between the groups at encoding when saccadic amplitude was measured, with participants in the Imagery First group producing smaller saccadic amplitudes than participants in the Recognition First group. One explanation for this could be that because participants in the Imagery First group were visualizing the pictures soon after they had seen them (and thus the pictures would still be in working memory), they may have focused on the main areas of interest. Whereas the Recognition First group had to remember a lot of pictures all at once (which would not be readily available in working memory) so scanned more widely to try and encode spatial relations between objects.

The lower average fixation duration at imagery for the Imagery First group compared to the Recognition First group suggests that visualizing a scene directly after you have seen it (Imagery First) is less cognitively demanding than visualizing it after the recognition task (Recognition First), where you have to choose from a number of inspected scenes. The lower average fixation duration at delayed imagery for the Imagery First group compared to the Recognition First group suggests that visualizing the scene after the recognition task makes it more cognitively demanding to visualize it again 48 hours later. In accordance with the Visual Buffer model, when you visualize the scene directly after inspection (Imagery First) this process facilitates the long term memory representation of the image, and thus makes it less demanding to visualize it a second time at a later occasion. It is possible that imagining a scene after recognition, where you have to choose from a number of pictures is a process that takes more cognitive processing than the visualizing per se, and therefore this does not facilitate the long term representation, and consequently makes it harder to visualize it a second time.

At recognition, participants in The Imagery First group may have made shorter fixations because they had ‘inspected’ each picture twice before the recognition test (once during encoding and once during imagery) so recognition may have been easier and less time at each fixation was needed.

Analyses of the number of fixations also varied according to the task being performed. There were more fixations at encoding and at recognition than at imagery or delayed imagery. Considering the above explanations of fixation duration, this makes sense because participants tried to take in as much at encoding as possible, making a greater number of shorter fixations. Research has shown that eye movements at encoding and recognition are similar (e.g. Foulsham & Underwood, 2008; Humphrey & Underwood 2008) and the current results support this, in that the numbers of fixations in these conditions are also similar. In the imagery conditions on the other hand, the longer fixation durations and greater saccadic amplitudes due to the more difficult task of recall with no visual cues may have ultimately lead to a smaller number of fixations in these conditions. This could also be due to the fact that there is less information to fixate on in a "mental image", and also because of reorganizing and resizing shown to occur during imagery. Some previous studies have shown a ‘shrinking’ of the mental image, (e.g. Finke, 1989; Johansson et al, 2006), though the saccadic amplitude results of this study suggest that a ‘stretching’ during imagery may also exist.

At recognition, average saccadic amplitudes in the Imagery First group were shorter than those in the Rec-

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

ognition First group. Taking into account the shorter fixation durations and increased number of fixations, this saccadic amplitude data suggests that participants in the Imagery First group focus on a smaller area of the picture. This could be because the participants in this group had, in effect, moved their eyes around the pictures twice before the recognition test – once at encoding and once at imagery, and thus were more familiar with where the areas of interest were situated. They therefore did not have to scan the picture as broadly as participants in the Recognition group, who had only seen the pictures once before.

Overall, average saccadic amplitudes were greater at imagery and delayed imagery than at encoding. This could be explained by the reorganizing and re-shaping shown to occur during imagery. As mentioned above, previous research has indicated a 'shrinking' of the mental image during imagery tasks, whereas the saccadic amplitude data in this paper suggests enlarging or 'stretching' of the mental image. One possible explanation for this could be a type of boundary extension, which has been shown to occur during imagery as well as perception (e.g. Intraub, Gottesman, and Bills, 1998).

The fact that the results showed no reliable difference between the imagery and delayed imagery conditions suggests that the reorganizing of mental images may take place between encoding and first imagery and then stays relatively stable over multiple imagery tasks.

Scan patterns produced at each condition were compared to every other condition using string analysis to create a similarity score. In The Imagery First group, scan patterns were more similar when comparing imagery and delayed imagery than when comparing encoding and imagery or encoding and delayed imagery or imagery and recognition. This could be explained in terms of mixed and pure process comparisons. When comparing imagery and delayed imagery, the task was the same in The Recognition First group and very similar in The Imagery First group, in that both conditions involved recalling a memory without any immediate visual cues. This could be referred to as a 'pure process comparison'. Whereas when comparing encoding and imagery or encoding and delayed imagery or imagery and recognition, one of the conditions in each comparison involved visual input from the stimulus and the other involved recalling without any visual input. These could be referred to as 'mixed process comparisons', and produce lower similarity scores.

In The Recognition First group, scan patterns are less similar when comparing encoding and imagery than when comparing encoding and recognition, or imagery and recognition or imagery and delayed imagery. Encoding and imagery is a mixed process comparison and it makes sense that scan patterns in these two conditions would be less similar than when comparing encoding and recognition or imagery and delayed imagery, as these are pure process comparisons. How then can we explain why there is such great similarity between imagery and recognition in The Recognition First group when this is a mixed task comparison, and the same result is not true of this comparison in The Imagery First group? In The Imagery First group, participants visualized the picture shortly after seeing it; therefore the visual image was still in short term memory and imagery involved more reconstruction of the picture rather than retrieval of the memory. It could be said that the spatial information was still in the visual buffer. In The Recognition First group, retrieval was a more competitive process due to the distracter stimuli in the recognition test. Participants had to remember which picture the description was referring to before imagining specific details or features, so this type of imagining is more like the process of recognition. It could be argued that the visual information had to be retrieved from long term memory and re-created in the visual buffer before the picture could be imagined. This also applies to the delayed imagery test and explains the high similarity in between recognition and delayed imagery in both Tasks. In this sense, the comparison between imagery with written cues and recognition is more of a pure process comparison than between encoding and imagery or encoding and recognition or encoding and delayed. The reproduction of eye movements at imagery argues against a purely bottom-up explanation of scan pattern similarity, as there is no visual (bottom-up) information at imagery.

The most similar scan patterns came from pure process comparisons where there was similar visual input in each condition (imagery compared to delayed imagery and encoding compared to recognition), and from comparisons that mimicked the same retrieval processes (imagery compared to recognition in The Recognition First group and delayed imagery compared to recognition in both Tasks 1 and 2). Pure process comparisons could also offer an explanation for the similarities between encoding and recognition phases with regards to fixation durations and number of fixations. The lowest scan pattern similar-

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

ity scores came from mixed process comparisons (encoding compared to imagery, encoding compared to delayed imagery, and imagery compared to recognition in The Imagery First group).

Even though the string similarity scores were quite low when comparing encoding and imagery, (Imagery First group = 0.170; Recognition First group = 0.165), the scores were still reliably above chance, suggesting that eye movements are still reproduced even when no visual information is present (during imagery). This argues against a purely bottom-up explanation of scan pattern similarity.

The lower scan pattern similarity scores when comparing encoding and imagery could be due to reorganizing and re-sizing during mental imagery. However, the greatly increased similarity scores when comparing imagery and delayed imagery (Imagery First group = 0.274; Recognition First group = 0.346) suggest that reorganisation occurs mostly between encoding and imagery but then stays relatively stable over multiple imagery tasks.

Overall, the scan pattern analyses have shown that the more similar the retrieval process is to the encoding process, the more similar the scan patterns produced. This suggests that the visual buffer model may be more complicated than simply shifting attention to different parts of an internal image (Kosslyn, 1994). The relationship between the encoding and retrieval process seems to be very important and one might even suggest the existence of facilitatory and inhibitory pathways within the model. For example, retrieval of a representation from long term memory could be facilitated if exactly the same visual information is present at encoding and recognition, as there are more visual guides and less chance of reorganizing or resizing as the information is transferred from long term memory to the Visual Buffer. The cognitive load on working memory is also lowered.

Propositional accounts such as that of Pylyshyn (2002) argue that there is no such thing as a visual buffer and that when participants are asked to "imagine X" they use their knowledge of what "seeing X" would be like, and they simulate as many of these effects as they can. However, it seems very unlikely that participants are able to mimic behaviour so precisely in their eye movements. In agreement with Johansson et al (2006), the number of points and the precision of the eye movements to each point are too high to be remembered without a support to tie them together in a context, such as an internal image. This is backed up further by the finding that temporal information as well as spatial information is reproduced at retrieval and is consistent over time as long as the same retrieval process is used. Furthermore, if participants did store spatial scene information as a large collection of propositional statements, scan pattern similarity should have remained constant across conditions despite changing the retrieval task, but this was not the case.

The finding that scan patterns at imagery were highly similar to those at delayed imagery (48 hours later) suggests that they are stable over time. The similarity between the scan patterns also lends support for Perceptual Activation Theory, which suggests that since there is no actual diagram or picture to be seen during the imagery period, it is likely that an internalised cognitive perceptual model is in control of these scan patterns. PA theory states that perceptual experience consists in the ongoing activity of schema-guided perceptual exploration of the environment and that imagery is experienced when a schema that is not directly relevant to the exploration of the current environment is allowed at least partial control of the exploratory apparatus.

To conclude, in accordance with Johansson et al (2006), the results of this paper lend support for the visual buffer model of imagery (Kosslyn, 1994), and challenge the propositional visual index model (Pylyshyn, 2002). The variations in scan pattern similarities caused by manipulation of the retrieval processes suggests that the visual buffer may be more complicated than previously thought, with possible facilatatory and inhibitory pathways. The replication of scan patterns during imagery lends support for the Perceptual Activation Theory and argues against the fixation selection being based on low level factors. The lower scan pattern similarity scores when comparing encoding and imagery suggests that most of the re-sizing and reorganising of mental images occurs at this stage. The high scan pattern similarity scores when comparing imagery and delayed imagery suggests that much less resizing happens once the mental images have been formed and that these scan patterns are relatively stable over time.

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

References

Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience, 9*, 27-38.

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

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2):2, 1-19

Finke, R.A. (1989). Principles of Mental Imagery, Cambridge, MA: MIT Press.7

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8(2)*:6, 1-17.

Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: An fMRI study. *Cognitive Brain Research, 20*, 226-241.

Hacisalihzade, S. S., Allen, J. S., & Stark, L. (1992). Visual perception and sequences of eye movement fixations: A stochastic modelling approach. *IEEE Transactions on Systems Man And Cybernetics, 22*, 474-481

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends In Cognitive Sciences, 7*, 498-504.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. L. (2007). Visual saliency does not account for eye movements during search in real-world scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537-562). Oxford: Elsevier.

Holsanova, J., Hedberg, B., & Nilsson, N. (1998). Visual and verbal focus patterns when describing pictures. In W. Becker, H. Deubel & T. Mergner (eds.), *Current Oculomotor Research: Physiological and Psychological Aspects.* New York: Plenum (pp. 303-304).

Humphrey, K. & Underwood, G. (submitted). Domain knowledge moderates the influence of visual saliency in scene recognition.

Intraub, H., Gottesman, C. V., & Bills, A. J., (1998). Effects of perceiving and imagining scenes on memory for pictures. *Journal of experimental psychology. Learning, memory, and cognition, 24(1)*, 186-201.

Ishai, A., and Sagi, D. (1995). Common mechanisms of visual imagery and perception. *Science, 268*, 1772-1774

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40* (10-12), 1489-1506.

Johansson, R., Holsanova, J., & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science, 30*, 1053-1079.

Josephson, S., & Holmes, M. E. (2002). Attention to repeated images on the World-Wide Web: Another look at scanpath theory. *Behavior Research Methods, Instruments, & Computers, 34*, 539-548.

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

Kosslyn, S. M. (1994). *Image and Brain.* Cambridge, Mass.: The MIT Press.

Kosslyn, S. M., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995). Topographical representations of mental images in primary visual cortex. *Nature, 378*, 496-498.

Laeng, B., & Teodorescu, D. S. (2002). Eye scanpaths during visual imagery re-enact those of perception of the same visual scene. *Cognitive Science, 26*, 207-231.

Mast, F. W., & Kosslyn, S. M. (2002). Eye movements during visual mental imagery. *Trends in Cognitive Science, 6*, 271-272.

Neisser, U. (1976). *Cognition and reality*. San Francisco: Freeman.

Noton, D., & Stark, L. (1971a). Scanpaths in Saccadic Eye Movements While Viewing and Recognizing Patterns. *Vision Research, 11*, 929-942.

Noton, D., & Stark, L. (1971b). Eye movements and visual perception. *Scientific American, 224*, 34-43.

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692

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

Pieters, R., Rosbergen, & Wedel, M. (1999). Visual attention to repeated print advertising: A test of scanpath theory. *Journal of Marketing Research, 36.* 424-438.

Privitera, C. M. (2006). The scanpath theory: its definition and later developments. *Proceedings of SPIE (The International Society for Optical Engineering), 6057*.

Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, 970-982.

Privitera, C. M., Stark, L. W. & Zangemeister, W. H. (2007). Bonnard's representation of the perception of substance. *Journal of Eye Movement Research, 1(1):3,* 1-6.

Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in Cognitive Science, 4,* 197-207.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision, *Cognition*, 80 (1/2), 127-158.

Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory, *Behavioral and Brain Sciences*, 25 (2), 157-238.

Pylyshyn, Z. W. (2003). Return of the mental image: Are there really pictures in the brain? *Trends in Cognitive Science, 7,* 113-118.

Sheth, B. R. & Shimojo, S. (2001). Compression of space in visual memory. *Vision Research, 41,* 328-341.

Stark, L., & Ellis, S. R. (1981). Scanpaths revisited: cognitive models direct active looking. In D. F. Fisher, R. A. Monty & J. W. Senders (Eds.), *Eye movements: cognition and visual perception* (pp. 193-227). Hillsdale, NJ: Lawrence Erlbaum.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research, 45*(5), 643-659.

Tatler, B. W. (2007). The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14), 4, 1-17.

Thomas, N. J. T. (1999). Are Theories of Imagery Theories of Imagination? An Active Perception Approach to Conscious Mental Content. *Cognitive Science, 23*, 207-245.

Underwood, G., & Foulsham, T., (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, *59*, 1931 - 1949

Underwood, G., Foulsham, T., van Loon, E., Humphreys, L., & Bloyce, J. (2006). Eye movements during scene inspection: A test of the saliency map hypothesis. *European Journal of Cognitive Psychology*, *18*, 321-342.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-word scenes. *Consciousness & Cognition, 17(1),* 159-170.

DOI 10.16910/jemr.2.2.3

ISSN 1995-8692